Item-Level Heterogeneity in Value Added Models: Implications for Reliability, Cross-Study Comparability, and Effect Sizes

Joshua B. Gilbert
Harvard University

Zachary Himmelsbach
Harvard University

Luke W. Miratrix
Harvard University

Andrew D. Ho
Harvard University

Benjamin W. Domingue
Stanford University

Value added models (VAMs) attempt to estimate the causal effects of teachers and schools on student test scores. We apply Generalizability Theory to show how estimated VA effects depend upon the selection of test items. Standard VAMs estimate causal effects on the items that are included on the test. Generalizability demands consideration of how estimates would differ had the test included alternative items. We introduce a model that estimates the magnitude of item-by-teacher/school variance accurately, revealing that standard VAMs overstate reliability and overestimate differences between units. Using 16 academic outcomes from 8 studies with item-level data, we show how standard VAMs overstate reliability by an average of .09 on the 0-1 reliability scale (median = .04, SD = .10) and provide standard deviations of teacher/school effects that are on average 12% too large (median = 3%, SD = 23% points). We discuss how imprecision due to heterogeneous VA effects across items attenuates effect sizes, obfuscates comparisons across studies, and causes instability over time, though these effects are attenuated when the number of items is high. Our results suggest that accurate estimation and interpretation of VAMs requires item-level data, including qualitative data about how items represent the content domain.

VERSION: August 2025

Item-Level Heterogeneity in Value Added Models: Implications for Reliability, Cross-Study Comparability, and Effect Sizes

Joshua B. Gilbert[1], Zachary Himmelsbach[1], Luke W. Miratrix[1], Andrew D. Ho[1], and Benjamin W. Domingue[2]

[1]Harvard Graduate School of Education
[2]Stanford Graduate School of Education

August 15, 2025

Abstract

Value added models (VAMs) attempt to estimate the causal effects of teachers and schools on student test scores. We apply Generalizability Theory to show how estimated VA effects depend upon the selection of test items. Standard VAMs estimate causal effects on the items that are included on the test. Generalizability demands consideration of how estimates would differ had the test included alternative items. We introduce a model that estimates the magnitude of item-by-teacher/school variance accurately, revealing that standard VAMs overstate reliability and overestimate differences between units. Using 16 academic outcomes from 8 studies with item-level data, we show how standard VAMs overstate reliability by an average of .09 on the 0-1 reliability scale (median = .04, SD = .10) and provide standard deviations of teacher/school effects that are on average 12% too large (median = 3%, SD = 23% points). We discuss how imprecision due to heterogeneous VA effects across items attenuates effect sizes, obfuscates comparisons across studies, and causes instability over time, though these effects are attenuated when the number of items is high. Our results suggest that accurate estimation and interpretation of VAMs requires item-level data, including qualitative data about how items represent the content domain.

Keywords: value-added model, generalizability theory, reliability, education policy, accountability

Corresponding author: joshua_gilbert@g.harvard.edu

The authors wish to thank Mike Hardy, Doug Mosher, Eric Taylor, Alex Bolves, Lily An, seminar participants at Harvard University, Stanford University, and the International Meeting of the Psychometric Society, and three anonymous reviewers for their helpful comments on drafts of this paper. This research was supported in part by the Jacobs Foundation (BD). This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D240025 to

Stanford University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education (BD). The authors report no conflicts of interest.

Author Contributions: Conceptualization: JG, AH; Methodology: JG; Software: JG; Formal Analysis: JG, ZH; Writing—original draft preparation: JG; Writing—review and editing: All.

Data and Code Availability: The RCT datasets are available in the Item Response Warehouse (IRW, Domingue et al., 2024) under the prefix `gilbert_meta`: https://redivis.com/datasets/as2e-cv7jb41fd. Our data, code, results, and supplemental materials will be available at the following URL upon publication: https://doi.org/10.7910/DVN/89YITQ.

1 Introduction

Value-added models (VAMs) of teacher and school effects play an important role in education research as they promise to provide estimates of teacher and school effectiveness that account for the non-random sorting of students into classrooms or schools (Chetty, Friedman, & Rockoff, 2014a; Harris, 2009). Research demonstrates that VA estimates are more predictive of teacher and school quality than alternative metrics such as teacher credentials and also that VA varies widely across teachers and schools (Aaronson et al., 2007; Goldhaber et al., 2013; Rivkin et al., 2005). As a result, VA estimates are commonly used as both predictors of future student outcomes (e.g., high school graduation, income), and as outcomes in themselves to determine what observable features (e.g., teacher years of experience, school size) predict VA to better understand the contribution of teachers and schools to student outcomes (Aslantas, 2020; Chetty, Friedman, & Rockoff, 2014b; Cowan et al., 2023; Hanushek & Rivkin, 2010). While generally applied to student achievement in math and language, VAMs are flexible and have also been applied to alternative outcomes such as social-emotional learning or attendance (Jackson, 2018; Jackson et al., 2020; Liu & Loeb, 2021) and other distributional features of test scores such as within-school student variances (Leckie et al., 2024). Given the prevalence of VA research in education, the statistical properties, methodological considerations, and policy implications of VAMs have been the subject of extensive discussion and debate over the past 40 years (American Educational Research Association, 2015; Amrein-Beardsley et al., 2016; Bacher-Hicks & Koedel, 2023; Cawley et al., 1999; Chetty, Friedman, & Rockoff, 2014; Everson, 2017; Koedel et al., 2015; Levy et al., 2019; Manzi et al., 2014; Morganstein & Wasserstein, 2014; Page et al., 2024; Pivovarova et al., 2016; S. Raudenbush & Bryk, 1986; S. W. Raudenbush, 2004; Schochet & Chiang, 2013).

Beyond their use in empirical research, state accountability systems have put VAMs to practical use to identify—and subsequently reward or punish—highly effective and less effective teachers (Konstantopoulos, 2014). In the United States, for example, under Race to the Top, VAM measures were a required component of states' accountability systems. Since the passage of the Every Student

Succeeds Act (ESSA), the use of VAMs has declined, but as of 2018, 15 states still used VAMs in their accountability systems (Close et al., 2018). In some cases, VAMs have informed high-stakes decisions: the DC Impact program used VAMs to identify "minimally effective" teachers and fire them if they did not improve within one year (Dee & Wyckoff, 2015). Similarly, Hanushek (2011) suggests that replacing the least effective teachers, as measured by VAMs, with average teachers would increase student welfare dramatically. Additional states use VAMs for lower-stakes purposes.

Common methodological questions related to VAM include identification and reliability. That is, (1) to what extent do VA estimates provide unbiased *causal* impacts of teachers or schools on student performance, and (2) how *reliable* are VA estimates of individual teachers and schools? The former question of causal identification in VAMs has received extensive commentary in the literature (J. Angrist et al., 2024; J. D. Angrist et al., 2017; Bitler et al., 2021; Kane et al., 2013; Koedel et al., 2015; Reardon & Raudenbush, 2009; Rothstein, 2010; Rubin et al., 2004); the latter question of VA reliability motivates the present study. Questions of causal identification aside, issues of VA reliability are critical because the appropriate use of VA estimates in practice is often contingent on the precision of the estimate. For example, if VA estimates are imprecise, then teachers or schools could be arbitrarily punished or rewarded in ways that do not reflect differences in their true underlying effectiveness (Amrein-Beardsley, 2014), judgments of differences in student growth rates would be incorrect (Lockwood & Castellano, 2015; Monroe & Cai, 2015; Wells & Sireci, 2020), and therefore the incentive effects of VA-based accountability structures would be weakened (Brehm et al., 2017).

In this study, we consider that student outcomes used in VAMs are typically aggregates of test items and therefore a teacher or school's contribution to student learning may vary across the individual items of an outcome measure. In other words, we investigate whether the impact of a teacher or school on a student's tendency to answer a given test item correctly can differ markedly from the teacher or school's impact on other items in the same test. Classic VAMs implicitly estimate the average impact of the teacher or school (henceforth "cluster" to maintain generality) over the set of items used, and thus do not take the representativeness of the items themselves into account when estimating uncertainty of the VA estimates (e.g., Koedel et al., 2015). This is an important concern if tests contain different items from year to year.

We show analytically and verify through simulation that when such cluster-by-item interactions are ignored, estimates of both VA reliability and the variation in teacher or school effectiveness can be upwardly biased, sometimes markedly so, thus inflating apparent differences between clusters. We then apply our approach to 16 academic outcomes across 8 empirical studies in education with item-level outcome data and baseline test scores conducted in the United States. We find that cluster-by-item interactions are both prevalent and large in magnitude, leading to an average overestimation of VA reliability by .09 (SD = .10) on the 0-1 reliability scale and provide standard deviations of VA effects that are on average 12% too large (SD = 23% points). Thus, researchers using standard approaches to VAM may be overestimating the reliability and variability of VA when test items vary across test administrations (or when the intention is to draw inferences regarding a larger pool of possible items). As VAMs are typically based on total test scores, rather than item-level scores, this item-level variation and the subsequent reduction in VA reliability is often obscured.

The study is organized as follows. We review the rationale for VAMs and standard approaches to estimating VAMs in Section 1.1. We discuss standard methods for estimating the reliability of VA estimates in Section 1.2 and extend VAMs to individual item responses in Section 1.3. We outline our methods and empirical data in Section 2. We examine results in Section 3. Section 4 concludes with a discussion of policy implications, limitations, and future directions.

1.1 Value-Added Models (VAMs)

A standard approach to VAMs is to model student performance, typically represented by a year-end achievement test score, as a function of membership in cluster k, controlling for baseline scores:

$$\text{post}_{jk} = \beta_0 + \beta_1 \text{pre}_{jk} + u_k + \theta_{jk} \tag{1}$$

$$u_k \sim N(0, \sigma_u^2) \tag{2}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2). \tag{3}$$

Here, post_{jk} and pre_{jk} are the year-end and prior-year test scores for student j in cluster k, β_0 is mean student performance when $\text{pre}_{jk} = 0$, u_k is the cluster effect, and θ_{jk} is the student residual. u_k represents the residual cluster effect on posttest scores that is not accounted for by pretest scores. VAMs often include covariates beyond pretest scores such as demographic variables to adjust for other forms of student sorting within clusters that would otherwise bias estimates of cluster effects (Levy et al., 2023). In other words, VAMs estimate the aggregate conditional status of students in a cluster given the covariates (Castellano & Ho, 2015), and a causal interpretation of the effect of u_k on student performance is justified to the extent that the observed covariates capture relevant pre-existing differences between clusters in terms of both the growth rate of the students as well as true baseline ability. That is, the identification strategy underlying a VAM is a selection on observables framework (Bacher-Hicks & Koedel, 2023; Rothstein, 2009).

While many alternative approaches to VAMs are available, such as cluster fixed effects, student fixed effects, two-step approaches, multiple pretests, gain scores rather than covariate adjustment, jackknife approaches, and others (see Koedel et al., 2015 and Bacher-Hicks and Koedel, 2023 for reviews), we use the simple framework of Equation 1 throughout this study for clarity of exposition (we consider alternative frameworks in Section 4). Furthermore, although cluster fixed effects approaches are perhaps the most common VAM estimation strategy in practice, we use a cluster random effects approach because we need the variance of u_k (σ_u^2) to calculate reliability, and the random effects model provides a consistent estimate of this variance. While random effects models assume normal distributions on the random effects terms, model results tend to be robust to violations of this assumption (Bell et al., 2019; Schielzeth et al., 2020).

1.2 Reliability of VA Estimates

Critically, u_k is unobserved. While u_k can be estimated from a statistical model by averaging the student residuals in each cluster or with fixed effects or with empirical Bayes shrinkage estimators, the estimate will contain measurement error that must be accounted for in subsequent analyses to avoid bias (Lockwood & McCaffrey, 2020; McCaffrey et al., 2009). To quantify the degree of measurement error in an estimate of u_k, we can estimate its reliability. Reliability is defined as the ratio of true score variance to observed score variance. More formally, in a Classical Test Theory framework (Lord & Novick, 1968), we can decompose an observed score X into the sum of a true score T and random measurement error E: $X = T + E$. Under the assumption that the the error term is independent of the true scores (i.e., $E \perp\!\!\!\perp T$), we can decompose the variances as follows: $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Reliability, denoted ρ, is therefore defined as $\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$. Reliability values range from 0 to 1, where 0 indicates that observed variation is random noise and 1 indicates that all observed variation reflects persistent underlying variation. An equivalent interpretation is that reliability is the expected correlation between scores over replications, where 0 indicates no correlation between replications and 1 indicates perfect correlation between replications. In general, when using estimated scores in a second-stage analysis, lower reliability attenuates correlations between variables and reduces statistical power (Kline, 2023; Revelle & Condon, 2019).

Adapting the Classical Test Theory conception of reliability to the VAM case is straightforward. An observed VA estimate (i.e., the average student residual in each cluster) is equal to its true value u_k plus the realized mean of the student residuals for cluster k, denoted $\overline{\theta_{.k}}$. Assuming homoskedasticity, the variance of $\overline{\theta_{.k}}$ is $\frac{\sigma_\theta^2}{J_k}$, where J_k is the number of students in cluster k. Assuming a constant cluster size $J_k = J$ for simplicity, the observed variance of VA estimates is $\sigma_u^2 + \frac{\sigma_\theta^2}{J}$ because u_k and $\overline{\theta_{.k}}$ are independent. Applying the Classical Test Theory reliability formula to these values yields:

$$\rho_k = \frac{\sigma_T^2}{\sigma_X^2} = \frac{V(u_k)}{V(u_k) + V(\overline{\theta_{.k}})} = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\theta^2}{J}}. \tag{4}$$

In other words, Equation 4 provides the ratio of true cluster variance (σ_u^2) to the variance in estimated VA scores ($\sigma_u^2 + \frac{\sigma_\theta^2}{J}$), matching the Classical Test Theory formulation. Furthermore, Equation 4 can easily be extended to additional levels of hierarchy or other facets of variation, such as occasions, raters, or items as desired in a Generalizability Theory framework (Brennan, 2001). The ratio expressed by Equation 4 is mathematically equivalent to the expected correlation between VA scores over replications. An advantage of Equation 4 is that it is estimable with data from only one replication, making it attractive when calculating correlations between replications directly is impractical or impossible. As such, Equation 4 and its extensions are common in practice when estimating the reliability of cluster scores, in VAM contexts or otherwise (e.g., Jeon et al., 2009).

1.3 Incorporating Item-Level Data into VAMs

In most empirical applications, the posttest score is a single-number scaled score, constructed from student responses to individual assessment items. When the item responses are available, we can add a level to the model, in which items are indexed by i and the student j in cluster k's response to item i, y_{ijk}, is modeled directly:

$$y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + u_k + \theta_{jk} + b_i + e_{ijk} \tag{5}$$

$$u_k \sim N(0, \sigma_u^2) \tag{6}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{7}$$

$$b_i \sim N(0, \sigma_b^2) \tag{8}$$

$$e_{ijk} \sim N(0, \sigma_e^2). \tag{9}$$

The key additions to this model include a random effect for item, b_i, that accounts for systematic variation in item easiness, and an error term e_{ijk}, capturing unexplained variability within students. u_k continues to represent VA on student performance, averaged across items. We consider y_{ijk} to be continuous for clarity of exposition but note that our arguments and results hold for dichotomous items that are more common in educational research. In our formulation, students are nested within clusters but crossed with items. That is, every student responds to every item, but a student is a member of only one cluster. Such designs are common, for example, when students across multiple classrooms or schools take the same standardized test. Considering both the set of students and items as random draws from a population, under Equation 5, the reliability of u_k is as follows, where I is the number of items:

$$\rho_k = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{10}$$

We do not include the item variance σ_b^2 in this equation because, so long as all students answer the same items, relative performance is not affected. That is, any variation in the average item difficulty on realizations of a test will shift the entire distribution up or down, but will not change the rank order of the respondents. Differences between VA reliability estimated with Equation 1 and Equation 5 will typically be negligible because $\frac{\sigma_e^2}{I}$ is absorbed by σ_θ^2 in Equation 1 (see also Appendix B). While Equation 1 is common in practice, we proceed with Equation 5 as our baseline for comparison to more clearly demonstrate the implications of cluster-by-item interactions for reliability.

Importantly for our purposes, Equation 10 assumes that the VA effects u_k are constant across items. This need not be the case: clusters may differentially add value to specific test items, above and beyond any average effect represented by u_k. Such heterogeneity of item-level effects is well-documented in randomized controlled trials, wherein treatment impacts may vary markedly across the items of the outcome measure (Ahmed et al., 2024; Gilbert, Himmelsbach, et al., 2025; Halpin & Gilbert, 2024). We can allow for VA effects to vary by item by adding an interaction term to the model, in which ν_{ik} represents the residual VA effect on item i by cluster k after the main effect u_k has been accounted for:

$$y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk} \tag{11}$$

$$u_k \sim N(0, \sigma_u^2) \tag{12}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{13}$$

$$b_i \sim N(0, \sigma_b^2) \tag{14}$$

$$\nu_{ik} \sim N(0, \sigma_\nu^2) \tag{15}$$

$$e_{ijk} \sim N(0, \sigma_e^2). \tag{16}$$

The variance of these interactions, σ_ν^2, captures the variability of VA effects at the item level. As a concrete example, a positive value of u_k indicates that cluster k improves student performance on average, across all items. A positive value of ν_{ik} implies that, net of any average effect, cluster k further improves student performance on item i. The total cluster effect on item i is $u_k + \nu_{ik}$.

Why might cluster effects vary across individual test items? One reason could be accountability structures that incentivize teachers or schools to differentially focus on content within a test. For example, Jacob (2005) shows that improvements on basic math skills were larger than those on complex math skills following the introduction of test-based incentives. If the basic math skills are easier to improve, these results may reflect reallocation of teacher effort (Taylor, 2023), and may be related to issues of score inflation (Koretz, 2005, 2008), whereby improvements on item performance do not reflect improvements on the underlying trait being measured. Another explanation could be variation in within-teacher skills (Papay et al., 2020). For example, a teacher may simply be better at teaching proportions than geometry, and therefore their VA may be higher on proportion items compared to geometry items on a math test that includes both types of items. From the item perspective, some items may simply be more "instructionally sensitive" than others (Naumann et al., 2014; Polikoff, 2010). Whatever their causes, such effects would be captured by ν_{ik}.

The inclusion of ν_{ik} in the model implies an additional source of variation that affects the reliability of u_k. This occurs because, to the extent that test items vary across replications, the specific items selected and the pattern of cluster-by-item interactions become part of the total variance in item performance. Assuming that the items are a random sample of some larger pool of potential items, under Equation 11, the reliability of u_k is as follows, where σ_b^2 is still omitted because overall item easiness does not affect relative student (or cluster) performance:

$$\rho_k = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{17}$$

When $\sigma_\nu^2 = 0$, Equation 17 reduces to Equation 10. However, the addition of $\frac{\sigma_\nu^2}{I}$ to the denominator means that the reliability of u_k will decrease to the extent that $\sigma_\nu^2 > 0$, even as the number of students per cluster goes to infinity. In other words, ρ_k will only asymptote to 1 as $I, J \to \infty$, not only as $J \to \infty$.

Another way to illustrate the conceptual difference between Equations 10 and 17 is to consider the interpretation of reliability as correlation between replications. That is, Equation 10 provides the expected correlation between VA estimates when only *students* vary between replications, whereas Equation 17 provides the expected correlation between VA estimates when both *students and items* vary between replications. Because the specific items on a given realization of a test are typically not of interest in themselves, but rather, as representative of some broader domain (De Boeck, 2008), we argue that the reliability captured by Equation 17 is likely to be more meaningful in most empirical contexts. In other words, neither reliability is intrinsically correct or incorrect. Rather, both equations provide an estimate of different conditional reliabilities that depend on what facets of variation the researcher considers to be fixed or variable across replications. While Equation 17 is extendable to additional sources of variation such as occasions, in this study, we consider single-occasion estimates of reliability.

The reliability of VA estimates has received much commentary in the education policy literature, but the potential use of item-level data in VAMs has received relatively little attention outside the psychometric literature in, for example, item difficulty modeling (Prowker & Camilli, 2007) and the effects of rapid guessing (Jensen et al., 2018). Similarly, consideration of interactions between facets of variation (e.g., clusters and items, students and items, items and time, etc.) is common in Generalizability Theory applications (Brennan, 2001; Jeon et al., 2009), but as of yet, such considerations are rarely applied to VAMs. The closest example of our proposed approach is Hawley et al. (2017), who use multiple test score outcomes in a latent variable formulation of VAM. However, they do not examine the cluster-by-test interactions that would be most analogous to our approach.

The present study is therefore motivated by three primary research questions (RQs):

RQ1. What are the consequences of omitting the cluster-by-item interactions ν_{ik} from the model on the estimated variance of the cluster effects σ_u^2 and the estimated reliability of the estimated cluster effect u_k?

RQ2. What are typical magnitudes of σ_ν^2 relative to σ_u^2 in empirical data in education?

RQ3. To what extent does the presence of cluster-by-item interactions ν_{ik} affect empirical estimates of the variation and reliability of cluster effects in empirical data in education?

We examine RQ1 through an analytic derivation and confirm the results via simulation. We examine RQ2 and RQ3 through the analysis 16 datasets in education that contain academic outcome item responses and baseline test scores.

2 Methods

2.1 Analytic Derivation

We demonstrate that when cluster-by-item interactions ν_{ik} are present in the data-generating process but omitted from the estimation model, both the estimated cluster variance $\widehat{\sigma}_u^2$ and the reliability of u_k, $\widehat{\rho}_k$, are upwardly biased. In Appendix A, we provide an analytic derivation of these facts under the simplifying assumptions that the data are balanced, students are randomly assigned to clusters, and there are no covariates in the model. In that case, $\mathbb{E}[\widehat{\sigma}_u^2] = \sigma_u^2 + \frac{\sigma_\nu^2}{I}$, thus inflating estimated differences between clusters, and the remainder is distributed among the components of the denominator of Equation 17. As a result, Equation 10 produces an upwardly biased estimate of reliability when Equation 17 is the true data-generating process. The bias is approximately equal to the following, where $\widehat{\rho}_k$ is the estimated reliability under Equation 10:

$$\mathbb{E}[\widehat{\rho}_k] - \rho_k \approx \frac{\frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{18}$$

The bias will only be zero when $\sigma_\nu^2 = 0$ or as $I \to \infty$.

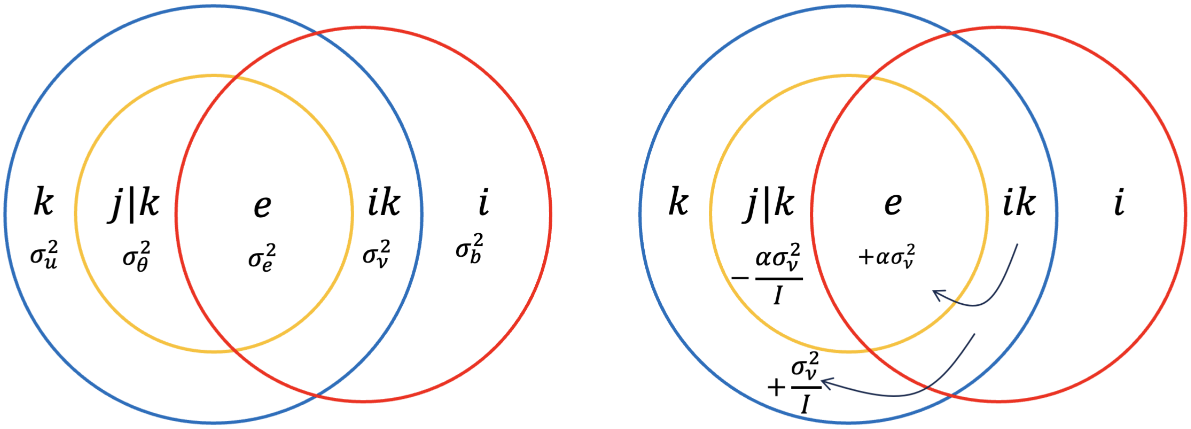
We can also understand these results on an intuitive level. Because the outcome is continuous, the total variance in y_{ijk} is constant across models and σ_ν^2 is distributed among the other variance components included in the model (Chan & Hedges, 2022; Lee & Hong, 2019; Shi et al., 2010; Ye & Daniel, 2017). Figure 1 summarizes how σ_ν^2 is absorbed by the other variance components in the model using a Venn diagram visualization (Brennan, 2001).

2.2 Simulation

Given the analytic derivation described in Appendix A, we examine model performance under more realistic assumptions than those required for the derivation in our simulation, including model misspecification, to determine how well the analytic result performs under more complex circumstances. Our data-generating model includes two additional parameters beyond Equation 11 to represent stratification common in educational systems: (1) a cluster covariate that predicts differences in cluster VA and (2) nonrandom sorting of students to clusters based on (here, perfectly reliable) pretest scores to match a common empirical justification for the use of VAM.

Table 1 summarizes the simulation design. The focal simulation factor is the proportion of item-level VA variance explained by cluster-by-item interactions, or $\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$. We fix the sum $\sigma_u^2+\sigma_\nu^2 = 1$ so

Figure 1: Illustrating Variance Components and Their Bias Under Misspecification



The left figure provides a schematic of the crossed and nested variance components (σ^2) from Equation 11, in which students j are nested within clusters k and crossed with items i. The right figure shows how omitted cluster-by-item interaction variance σ_ν^2 is absorbed by the remaining variance components in the model. I is the number of items, J is the number of students per cluster, and K is the number of clusters. $\alpha = \frac{JK-J}{JK-1}$. We ignore σ_b^2 because it is not necessary for the calculation of relative reliability. See Appendix A for additional detail.

that the total residual variance in y_{ijk} remains constant across conditions to facilitate comparability. When this proportion is 0, there are no cluster-by-item interactions, and the variance of the cluster effects is 1. When this proportion is 1, there are no average cluster effects, and the variance of the cluster-by-item interactions is 1. We vary this proportion from 0 to 1 in increments of .20. The other varying simulation factors include the number of items, the number of students, and the reliability of the pretest variable. We vary the number of outcome items at 5, 20, and 40 to represent short, moderate, and long assessments, the number of students at 500, 1000, and 2000 to represent small, moderate, and large sample sizes, and the pretest reliability at .75 and 1 to represent moderately and perfectly reliable tests. We fix the remaining values to those specified in Table 1 and perform 250 replicates across each of the 108 conditions.

We generate the data and fit two models to both a continuous and dichotomized outcome, one assuming constant VA effects across items, the other allowing for cluster-by-item interactions, resulting in four models in total.[1] Our primary goal is to examine estimates of the variance

[1] While we could fit logit models to the binary outcome variable, this would introduce interpretational complexities because misspecification of the random portion of a multilevel logit model creates known biases in the fixed and

Table 1: Fixed and Varying Simulation Design Factors

Simulation Factor	Notation	Values
Number of Subjects	JK	500, 1000, 2000
Number of Subjects Per Cluster	J/K	20
Number of Items	I	5, 20, 40
Total VA Residual Variance	$\sigma_u^2 + \sigma_\nu^2$	1
Prop. Total VA Variance Attributable to Items	$\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$	0, .2, .4, .6, .8, 1
Student Residual Variance	σ_θ^2	1
Pretest Coefficient	β_1	1
Cluster Covariate Coefficient	γ_1	.5
Intraclass Correlation on the Pretest	ICC	.25
Pretest Reliability	α_{pre}	.75, 1
Error Variance	σ_e^2	1
Item Variance	σ_b^2	1
Intercept	β_0	0

Notes: The pretest and cluster covariates are drawn from $N \sim (0, 1)$. We fix the ICC at .25 to represent moderate clustering typical in education (Hedges & Hedberg, 2007) and J/K at 20 to represent a typical US classroom size.

components and the reliability of u_k derived from each model and to determine how these estimates compare to our analytic results. We expect that the reliability results from the simulation will closely match those of the derivation because, while the simulation design is more complex due to the presence of covariates and non-random sorting of students into clusters, the variance components of the model are nonetheless independent conditional on the included covariates. We conduct the simulation and empirical analyses in R and use the `lme4` package to estimate the models (Bates et al., 2015). We include sample R code to fit the relevant models and some extensions in Appendix C.

random portions of the model. This occurs because the residual variance of the logit model is fixed at 3.29 for model identification (Austin & Merlo, 2017; Breen et al., 2018; Mood, 2010). Thus, we proceed with a linear probability model here to ease interpretability and comparability. Furthermore, linear models are more common in Generalizability Theory applications. We return to this issue in Sections 3.3 and 4.

2.3 Empirical Application

2.3.1 Data Sources

We apply our approach to datasets from a large collection of randomized controlled trials (RCTs) with item-level outcome and baseline score variables in a variety of fields (Domingue et al., 2024; Gilbert, Himmelsbach, et al., 2025). Here, we limit our analysis to the 16 outcomes from 8 studies that meet the following inclusion criteria to maximize similarity to common uses of VAM in the US:

1. The outcome measure must represent a common academic outcome such as reading, math, science, or social studies.

2. The study is conducted in the US.

3. Students are clustered in a higher-level unit such as classrooms or schools.

4. The data include a baseline measure, either a lagged outcome or a measure similar to the outcome.

Table 2 summarizes the datasets and shows a studies containing hundreds to thousands of students, dozens of items, students ranging from Kindergarten to Grade 5, and a range of outcomes including reading comprehension, vocabulary, other literacy skills, and math. When item-level data are available for the baseline measure, we construct scaled scores using a One-Parameter Logistic (1PL) IRT model. We examine only immediate post-intervention data any datasets with multiple follow-ups. While the original RCTs have some causal policy evaluation aim, we use these data to explore the reliability of VA estimates and the magnitude of cluster-by-item interactions in empirical data in US school contexts.

2.3.2 Empirical Models

We fit two models separately to each dataset: (1) item-level VAM assuming constant item effects, and (2) item-level VAM allowing for cluster-by-item interactions. Thus, the empirical models are

Table 2: Empirical Datasets

Dataset	N	K	I	$\frac{N}{K}$	Cluster	Baseline	Age	Outcome
1: Gilbert et al. (2023)	7797	110	30	70.88	school	MAP	G3	Reading Comprehension
2: Kim et al. (2023)	2174	30	20	72.47	school	MAP	G2	Reading Comprehension
7: Kim et al. (2024)	1352	30	36	45.07	school	MAP	G3	Vocabulary
8: Kim et al. (2024)	1303	30	29	43.43	school	MAP	G3	Reading Comprehension
11: Kim et al. (2021)	2565	30	24	85.50	school	MAP	G1	Vocabulary
12: Kim et al. (2021)	2580	30	24	86.00	school	MAP	G2	Vocabulary
21: Davenport et al. (2023)	3671	172	13	21.34	class	lagged	G5	Math
23: Bang et al. (2023)	886	41	38	21.61	class	lagged	K-G1	Math
76: Thai et al. (2022)	428	20	78	21.4	class	lagged	K	Math
77: Cabell et al. (2025)	1075	47	26	22.9	school	lagged	K	Language Fundamentals
78: Cabell et al. (2025)	1075	47	186	22.9	school	lagged	K	Vocabulary
79: Cabell et al. (2025)	1100	47	30	23.4	school	lagged	K	Narrative Language
80: Cabell et al. (2025)	1075	47	35	22.9	school	lagged	K	Vocabulary
81: Cabell et al. (2025)	1075	47	18	22.9	school	lagged	K	Science
82: Cabell et al. (2025)	1075	47	18	22.9	school	lagged	K	Social Studies
100: Gilbert, Domingue, and Kim (2025)	2118	30	12	70.60	school	MAP	G2	Vocabulary

Notes: N = number of students, K = number of clusters, I = number of items, G = grade. For additional information on these datasets, see Gilbert, Himmelsbach, et al. (2025). We include the original dataset IDs in our tables and figures to facilitate replicability and comparability with the source study. For baseline measures, "lagged" indicates that the baseline is identical to the outcome measure, and MAP is the Measure of Academic Progress reading assessment, a common standardized test in the US.

specified in reduced form as follows, in which T_k is the treatment indicator, with normal distributions on all random effects:

$$\text{Constant Item VAM: } y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + \gamma_1 T_k + u_k + \theta_{jk} + b_i + e_{ijk} \tag{19}$$

$$\text{Varying Item VAM: } y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + \gamma_1 T_k + u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk}. \tag{20}$$

We include the treatment indicator in the model because, to the extent that the treatment is effective in a cluster-randomized trial, the variation in u_k will increase by increasing the differences between treatment and control clusters at posttest.[2]

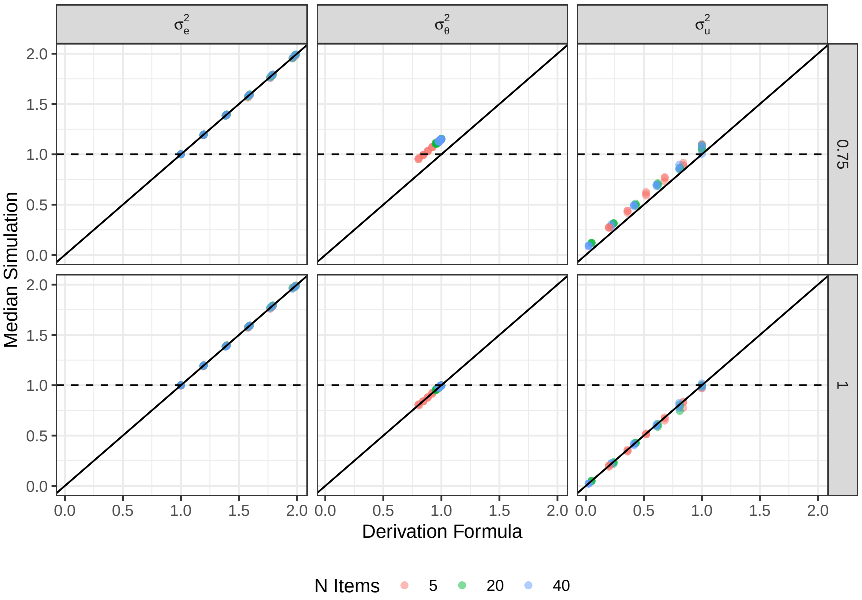
[2] We do not include the cluster mean of the pretest variable as an additional covariate in a Mundlak approach that relaxes the random effects assumption that the level-1 covariates are uncorrelated with the cluster effects (Antonakis et al., 2021; Mundlak, 1978; Rabe-Hesketh & Skrondal, 2022) because this can create bias in VAM contexts under non-random sorting of students to schools (Castellano et al., 2014; S. W. Raudenbush & Willms, 1995).

3 Results

3.1 Simulation

Figure 2 compares the analytic results of Appendix A to the estimated variance components from the simulation. We limit the figure to the continuous outcomes because the variance components for the dichotomous outcomes are on a different scale and are not expected to precisely conform to the analytic results. We see that in all but one case, the estimated variance components almost behave exactly as predicted. The exception is $\widehat{\sigma}^2_\theta$ when the pretest reliability is .75, in which case the observed variance is larger than predicted because the additional measurement error in the pretest is absorbed into the student variance.

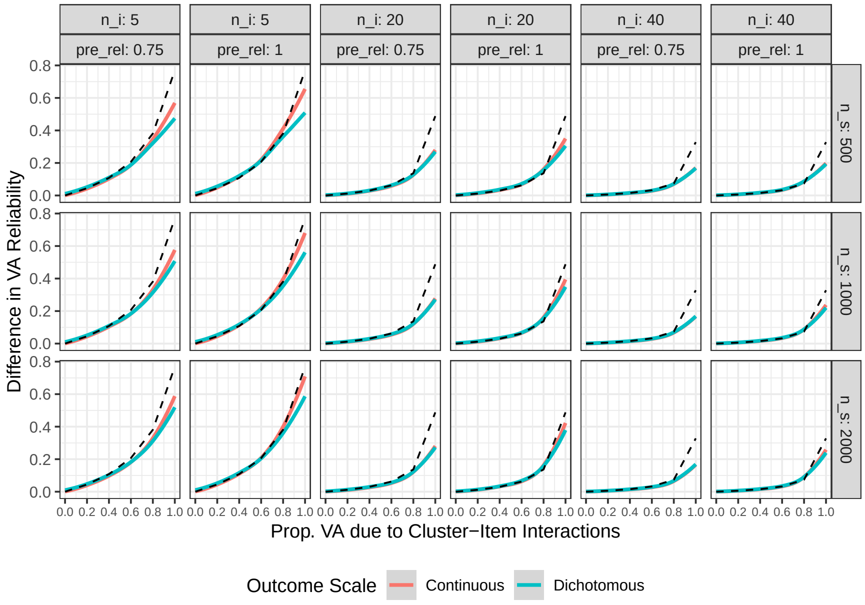
Figure 2: Comparison of Variance Components from the Simulations to the Predicted Values from the Analytic Derivation



The y-axis shows the median value of each variance component across simulations and the x-axis shows the value we would expect given the results in Appendix A. The plots are faceted by variance component (columns) and pretest reliability (rows). We omit σ^2_b from the figure because is it not needed in the reliability calculations.

Figure 3 compares the VA reliabilities of the simulation results to the analytic results derived in Appendix A. The x-axis shows the proportion of VA variance attributable to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$) and the y-axis shows the estimated difference in VA reliability (the misspecified model minus the correct model). The colored lines show LOESS curves fit to the simulation results for continuous and dichotomous outcomes, and the dashed black lines show the predicted bias derived from the formula in Appendix A based on each data-generating process. We first see that the simulation results are in close alignment with the analytic results that show that the estimated reliability of a constant VA model will be inflated whenever $\sigma_\nu^2 > 0$. Second, as expected, the bias is less severe when the number of items is greater. Last, the pattern of results is essentially equivalent regardless of the pretest reliability or whether the outcome is continuous or dichotomous. Thus, even in cases with relatively small person and item sample sizes and model misspecification due to dichotomization, the formula provided in Appendix A performs very well.

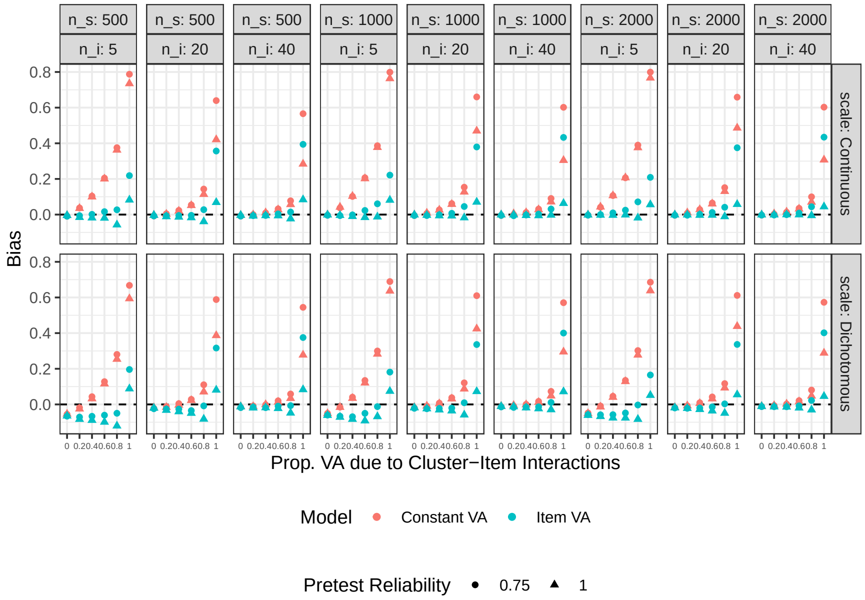
Figure 3: VAMs Assuming Constant Effects Across Items Overestimate Reliability



The y-axis shows the estimated difference in VA reliability (main effects model minus interaction model) and the x-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$). The black dashed lines represent the theoretical prediction of the reliability bias derived from the formula in Appendix A. The colored lines represent LOESS curves fit to the simulation results for continuous and dichotomous outcomes. n_i = number of items; n_s = number of students, pre_rel = pretest reliability.

Figure 4 shows the bias for reliability across the simulated conditions. We find that bias is near 0 for continuous outcomes and the appropriately specified model in all cases except when $\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2} = 1$, which occurs because $\widehat{\sigma}_u^2$ cannot be negative. In contrast, the constant VA model consistently provides an upwardly biased estimate of reliability whenever $\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2} > 0$. Accordingly, the RMSE for the appropriately specified model is consistently lower than that of the constant VA model for continuous outcomes. For dichotomous outcomes, the correctly specified model slightly understates the reliability of the latent continuous true score, an expected result given that the dichotomization significantly inflates σ_e^2 (Cohen, 1983), particularly when the number of items is low. When the number of items is high, the reliability bias under dichotomization is minimal, nearly matching the results for the continuous outcome.[3] Figure 5 shows the equivalent results for RMSE.

Figure 4: Reliability Bias by Simulation Condition

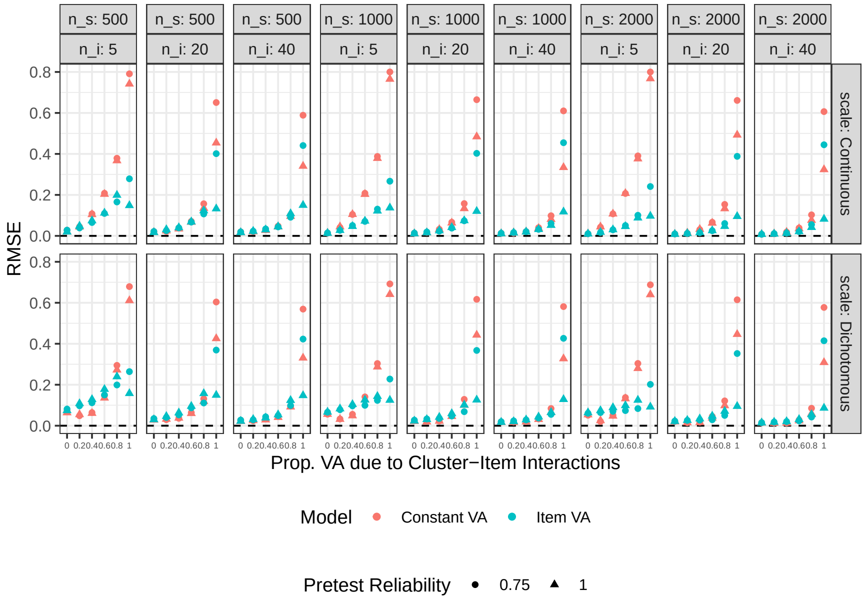


The y-axis shows the estimated bias in VA reliability and the x-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$). The points are color-coded by the analytic model and the shapes represent pretest reliability. n_s = number of subjects; n_i = number of items.

We include additional simulation results in our supplement. In short, we find that estimated standard errors for student and cluster covariates are unchanged, which is not surprising because the total variance remains constant across conditions. We also find that the coefficients for the

[3] Interestingly, the inflated $\widehat{\sigma}_e^2$ yields reliability estimates closer to the true value for the constant VA model for 5 dichotomous items when $\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$ is small but positive (e.g., .20) because these two effects work in opposite directions. See also Béland and Falk, 2022; Robitzsch, 2020 for an extended discussion of the performance of reliability metrics designed for continuous item responses with categorical data.

Figure 5: Reliability RMSE by Simulation Condition



The y-axis shows the estimated reliability RMSE and the x-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$). The points are color-coded by the analytic model and the shapes represent pretest reliability. n_s = number of subjects; n_i = number of items.

student and cluster fixed effects are unbiased across all conditions. Thus, the implications of omitted cluster-by-item interactions appear to only affect the reliability of VA estimates (by biasing the random effect variances) while the fixed portions of the model are relatively unaffected. The VA estimates themselves are perfectly correlated across the two models, though the lower reliability of the interaction model yields greater shrinkage in the associated VA estimates.

We also compare the item-level analyses to the more conventional VAM approach in which student mean scores rather than item responses serve as the outcome variable in the regression (Equation 1). We find identical results to those reported here. That is, as the proportion of VA variance due to cluster-by-item interactions increases, estimated reliability becomes upwardly biased. Thus, inflated reliability is not an artifact of mean scores *per se*, but of the omitted cluster-by-item interactions (see Appendix B). These interactions are masked in the standard mean score model but are easily estimable when item-level data are available.

3.2 Empirical Application

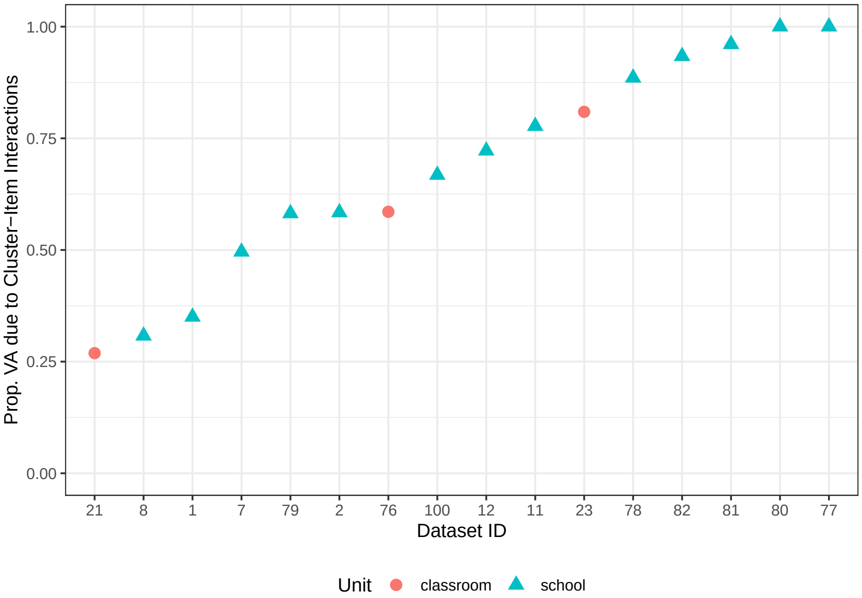
Our analytic sample of 16 academic outcomes from 8 studies contains 31,349 respondents (some of whom are included more than once because some studies include multiple outcome measures), 617 items, and 806,760 item responses. In general, outcome internal consistency is high (median = .85).

Figure 6 shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$) by dataset. We see that most datasets show extensive cluster-by-item variance, with most showing proportions over 50%. These results suggest that, far from a purely theoretical concern, cluster-by-item VA interactions are both large and prevalent in a wide range of empirical data. These findings are consistent with prior work on item-level heterogeneous treatment effects demonstrating that the effects of educational interventions often vary substantially across the items of the outcome measure (Ahmed et al., 2024; Gilbert, Himmelsbach, et al., 2025; Halpin & Gilbert, 2024). We include tables of the full model results for each dataset in our supplement.

Figure 7 shows the implications of these cluster-by-item interactions for the reliability of VA estimates. Because the number of students per cluster and the number of items answered by students varies in these data, we use the averages for each dataset in our reliability calculations. As expected from the analytic and simulation results, estimated VA reliability from the main effects model is in all cases equal to or greater than that of the interaction model (mean difference = .09, median = .04, SD = .10, on the 0-1 reliability scale). The differences in reliability are sometimes substantial, with 5 datasets showing inflation in estimated VA reliability of .10 or greater when cluster-by-item interactions are not accounted for. Accordingly, Figure 8 shows that the estimated SDs of the cluster effects ($\widehat{\sigma}_u$) derived from the main effects models are on average 12% greater than those of the interaction models (median = 3%, SD = 23% points). However, we emphasize that many datasets show small differences in reliability and $\widehat{\sigma}_u$ due to the relatively high number of items.

We include additional empirical results, including likelihood ratio tests comparing the models and profile confidence intervals for the variance components, in Appendix D. We find that the varying model is a significantly better fit to the data in all cases, and that the variance components are generally precisely estimated, particularly $\widehat{\sigma}_\nu$, due to the large number of cluster-by-item combinations in each dataset. Notably, the CIs for $\widehat{\sigma}_u$ overlap across models in all cases, suggesting that the proportional differences in point estimates shown in Figure 8 are small relative to their imprecision.

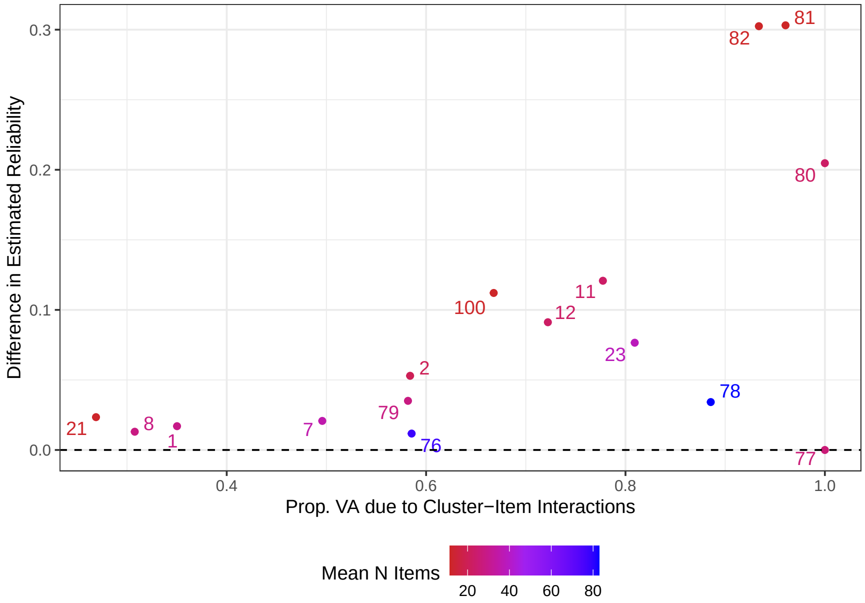
Figure 6: Proportion of Item-Level VA Variance Due to Cluster-by-Item Interactions in Empirical Datasets



The y-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma^2_\nu}{\sigma^2_u+\sigma^2_\nu}$) and the x-axis shows the dataset ID. Points are color-coded by the unit of clustering.

We replicate these analyses using IRT scaled scores instead of item responses as the outcome variable to match the more typical approach to VAM (Equation 1). We find essentially identical results to those reported here. Namely, using scaled scores rather than item responses as the outcome variable is essentially equivalent to the item-level model assuming constant VA effects (mean difference = .002, $p = .92$), whereas estimated reliability is significantly inflated compared to the cluster-item interaction model (mean difference = .087, $p < .001$). Thus, supporting the simulation results and analytic derivations, it is the omission of the cluster-item interactions from the model, not the construction of average or scaled scores *per se*, that leads to the most severe bias in estimated reliability.

Figure 7: Estimated Differences in VA Reliability in Empirical Datasets



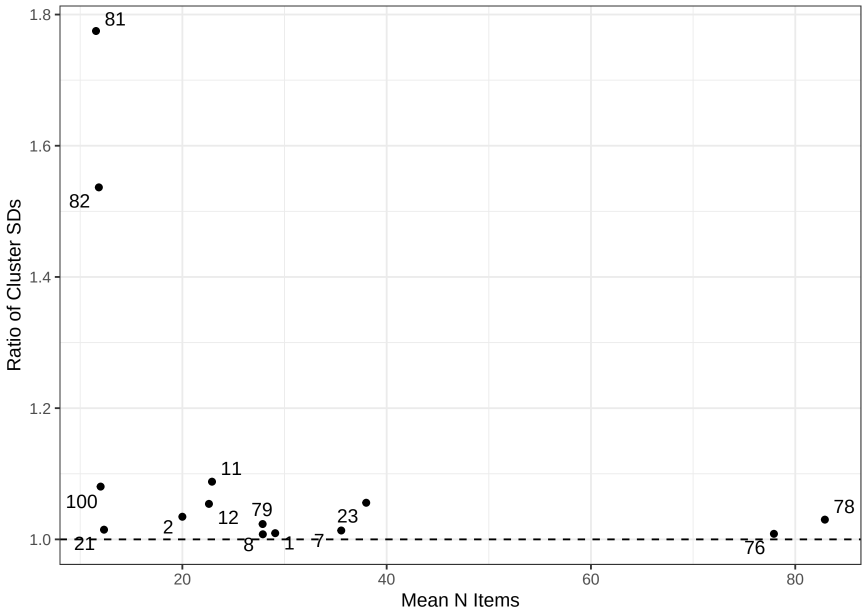
The x-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2}$) and the y-axis shows the estimated difference in reliability. The points are labeled by dataset ID. Points are color-coded by the mean number of items, which explains why the relationship is not strictly increasing. Dataset 77 has estimated reliability of 0 for both models.

3.3 Robustness Checks

Our empirical conclusions about inflated VA reliability and SDs of VA effects rest on the magnitude of σ_ν^2 relative to σ_u^2. Here, we consider three robustness checks to probe the sensitivity of our results, motivated by potential mechanisms that could affect σ_ν^2: (1) varying relationships between pretest scores and item scores, (2) non-linearity and ceiling/floor effects induced by the categorical item responses, and (3) inclusion of demographic covariates.

First, consider the assumption that the relationship β_1 between pre_{jk} and y_{ijk} is constant across all combinations of items and clusters. This assumption may not hold when, for example, a teacher notices relative strengths and weaknesses among their students and reallocates instructional effort accordingly. Such a mechanism would theoretically yield different relationships between pre_{jk} and y_{ijk} because, if a teacher focuses on a particular content area to compensate for student weaknesses, we might see a weaker relationship between pre_{jk} and y_{ijk} on items that measure

Figure 8: Estimated Ratio of VA SDs in Empirical Datasets



The x-axis shows the mean number of items answered by students and the y-axis shows the estimated inflation in $\widehat{\sigma}_u$ as a ratio (i.e., 2 means that the $\widehat{\sigma}_u$ is twice as large in the constant item effects model compared to the interaction model). The points are labeled by dataset ID.

those specific competencies. This would occur to the extent that added instruction in a content area reduces variation in student performance, thus weakening the relationship between pre_{jk} and y_{ijk} for relevant items. Similar reasoning applies if a teacher focuses on lower performing *students* to differentially improve performance on the types of items on which those students may struggle most. We explore this possibility by estimating a random slopes version of Equation 11 in which every cluster-item combination gets a unique slope for pre_{jk} (Donnellan et al., 2023). This could potentially reduce σ^2_ν. Applying this more flexible model to our empirical data shows similar but attenuated results to our primary analyses: the difference in VA reliability estimates between the two random intercepts and random slopes methods is on average about .01 and the mean proportion of item-level VA variance due to cluster-by-item interactions is 52%, compared to 68% for our main specification. Thus, differential prediction of item performance by pretest appears unlikely to fully account for our findings.

Second, all items from our empirical application are dichotomous rather than continuous. While some evidence suggests that VAMs are relatively robust to floor or ceiling effects (Koedel & Betts, 2010), such effects may be compounded when analyzing item-level data rather than aggregated test scores. For example, σ_ν^2 may be artificially inflated if items vary extensively in their difficulty, because a constant effect on overall student performance could manifest as variable effects on accuracy rates due to non-linear scaling, as in a logit model. That is, a constant improvement of one logit would bring an item with a baseline accuracy rate of 50% to 73% (23 percentage points), but would bring an item with a baseline accuracy rate of 88% to 95% (7 percentage points), creating the illusion of cluster-by-item interaction variance in a linear model when the true cause is non-linearity (Ho, 2008). We explore the consequences of dichotomous item responses by fitting cross-classified logit models to our empirical datasets because logit models can correct for ceiling and floor effects and related scaling issues (Domingue et al., 2022; Gilbert, Miratrix, et al., 2025). We find that the pattern of results is essentially unchanged from our main analyses, with a mean proportion of item-level VA variance due to cluster-by-item interactions of 65% compared to 68% in the linear models. Thus, the categorical item responses are unlikely to confound the large estimates of σ_ν^2 observed in the linear models.

Last, 12 datasets include demographic covariates, which we exclude from our primary models to maintain comparability across datasets. We rerun Equation 20 with all available covariates and compare the results to our main analysis. We find that the $\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$ is *higher* when controlling for covariates, at 76% on average in the covariate adjusted model compared to 71% in our primary specification. This finding suggests that the demographic covariates explain proportionally more between-cluster variation than cluster-by-item variation and further suggest that omitted demographic covariates are unlikely to be driving the large estimates of σ_ν^2 we observe. We list the available covariates in each dataset in Appendix E.

4 Discussion

4.1 Summary and Implications

VAMs have persisted as a standard method for evaluating teachers and schools in research and practice. While prior studies have examined the reliability of VA estimates (e.g., Briggs and Weeks, 2011; Konstantopoulos, 2014; Schochet and Chiang, 2013), the influence of the specific tested items has remained relatively unexplored. In this study, we show that cluster-by-item interactions are both large and prevalent in a wide range of empirical datasets in education. Thus, the implicit assumption of standard VAMs that cluster effects are constant across items appears unrealistic and leads to inflated estimates of VA reliability and differences between clusters. However, we emphasize that when the number of items is high, as is common in state accountability systems and many of our empirical examples, the biases may not be practically significant.

Our findings have several implications for the broader conversation about the use of VAMs in education research:

1. Cluster-by-item interaction variance exists in practice because some teachers and schools are better at improving some subskills on a test than others. This variance is large in empirical data. As a proportion of the total item-level VA variance, cluster-by-item interactions mostly exceed 50% in our sample of 16 academic outcomes from 8 studies in education in the US (Figure 6).

2. Ignoring cluster-by-item interaction variance in standard VAMs leads researchers to assume that reliability of cluster VA effects is higher than it is (Appendix A). The bias in reliability is zero only when the variance of cluster-by-item interactions is 0 or the number of items grows to infinity.

3. Researchers and policymakers who ignore cluster-by-item interaction variance may falsely conclude that their tests are sufficient to estimate VA at a desired level of reliability, when

in fact they may need longer tests. Thus, longer assessments may be necessary in high-stakes situations where high VA reliability is essential, especially when domains are broad and cluster-by-item interactions may be high. Such concerns are especially salient when practical guidance for VAM interpretation and use depends on heuristics such as minimum or "sufficient" sample sizes (e.g., American Educational Research Association, 2015, p. 450), rather than more nuanced statistical properties of VA estimates such as those explored here.

4. Measurement error due to cluster-by-item interactions attenuates correlational relationships between VA estimates and downstream outcomes, so these relationships would be stronger and perhaps more consistent if we correct for measurement error (Kline, 2023). Similarly, when used as outcomes, the measurement error in VA estimates will attenuate standardized effect sizes when estimated in two-step models (Gilbert, 2025). Thus, the predictive effects of VA on downstream outcomes and effects of predictors on VA may be underestimated while differences between clusters may be simultaneously overestimated.

5. Item-level data are necessary for a full accounting of VA effects. When available, researchers estimating VA reliability should use item-level data to account for potential cluster-by-item interactions. The degree of inflated reliability and variation of teacher and school effects in our empirical datasets would not be estimable from total scores alone. As shown in Figures 7 and 8, VA reliability and variation derived from standard models likely represents an upper bound on true reliability and variation. Thus, researchers should heed calls to share item-level data as part of their replication packages (Domingue et al., 2024).

4.2 Extensions and Future Directions

Given the large cluster-by-item interaction variances observed in the empirical data, what might explain such variance? We view this question as a promising area for future research. While Equation 11 assumes that the cluster-by-item effects ν_{ik} are idiosyncratic, it is nonetheless possible that ν_{ik} may reflect some shared influences omitted from the model, such as relative teacher proficiency on certain item clusters, as discussed in Section 1.3. By interacting cluster-level covariates with item-level covariates (e.g., teacher years of experience and whether an item is multiple choice or open response), our modeling framework easily allows researchers to specify and test hypotheses about the sources of cluster-by-item variance (see, e.g., Cohodes, 2016, who examines whether charter school impacts are consistent across subscales of a state test in an instrumental variables framework).

Similarly, item-specific VA estimates may serve as useful predictor variables if improvements on specific subskills within a domain are relevant for future student outcomes, or to identify a teacher's relative strengths and weaknesses for formative purposes (e.g., Papay et al., 2020). When qualitative data such as item text is also available, further research might then explore the extent to which different types of measures such as researcher-developed vs. independently-developed assessments show different degrees of cluster-by-item interactions (see Gilbert and Soland, 2024) or how qualitative evidence on the nature of the domain and how items represent the domain may help to understand differential patterns of VAM across items (Ho, 2024). We caution, however, against over-interpreting estimates of individual item-specific VA estimates, as these will tend to be imprecisely estimated unless the number of students per cluster is large. An individual cluster-by-item interaction ν_{ik} has reliability $\frac{\sigma^2_\nu}{\sigma^2_\nu + \frac{\sigma^2_\theta}{J} + \frac{\sigma^2_e}{J}}$.

A further promising area of extension would be to simultaneously consider multiple levels of hierarchy, such as students within teachers within schools. Whereas our empirical examples demonstrate 2-level structures, 3-level approaches could help determine to what extent between-school variance is explained by, for example, teacher-by-item interactions. Conversely, differences between teachers across schools may be partially explained by school-by-item interactions. A full decomposition of VA effects at the student, teacher, school, and item levels (and their interactions) could provide important insights into the sources of VA effects in a broad range of empirical contexts.

In a similar vein, we focus in this study on reliability estimates derived from a single measurement occasion. The extent to which cluster-by-item interactions may relate to changes in VA over time are complex and worthy of further study. It may be, for example, that some proportion of year-to-year variation in estimated VA reflects simultaneous changes in items across test administrations. Researchers and practitioners may conclude that VA estimates vary substantially from year to year and that they need to average over multiple years for stable estimates, when in fact this is in part due to cluster-by-item interaction variance, and they could also address this issue with longer tests, given that standardized test items typically shift from year to year (Holland & Dorans, 2006; Kolen & Brennan, 2014). More complex models that include clusters, students, items, occasions, and their interactions are a promising area of future research that would serve to further disentangle the interplay among these facets of variation.

Furthermore, our models are homoskedastic in that they assume that the cluster-by-item variance is constant across clusters. This need not be true, as some clusters may have relatively consistent impacts on item performance while others have more variable impacts. Extensions of our modeling approach to allow for heteroskedasticity, and explorations of the extent to which more or less consistent VA effects across items are themselves predictors of other student outcomes offer another promising avenue of exploration (Cárdenas-Hurtado et al., 2025; Leckie et al., 2024; Wiedermann et al., 2024).

The extent to which cluster-by-item interactions may complicate item selection in computer adaptive testing (CAT, Meijer and Nering, 1999) is another promising area of future research. In CAT, items are selected during testing to maximize the precision of an individual student's ability estimate (Bock & Mislevy, 1982). Cluster-by-item interactions can be interpreted as a type of differential item functioning whereby certain items are easier of harder for students in different clusters, conditional on the students' overall ability level, and may undermine standard item selection protocols that assume invariant item parameters. Thus, standard CAT procedures may have the potential for increased efficiency to the extent that stable cluster-by-item interactions are identified in advance and can be integrated into the CAT algorithms.

Last, we use a simple VAM formulation throughout this study for clarity of exposition, but as described in Section 1, many alternative approaches are possible, such as cluster fixed effects, gain scores, and others. For example, the popular procedure proposed by Chetty, Friedman, and Rockoff (2014a) uses a jackknife approach in which a VA estimate from a single year is regressed on VA estimates from all other years, and the prediction from this regression is used as the VA estimate. By using multiple years of data, the standard errors provided by the jackknife approach implicitly account for year-to-year variation in test items by attenuating the correlation between cross-year VA estimates and increasing residual variance, as described in Section 1.3. The jackknife approach is less relevant for our purposes because it is designed to provide unbiased prediction of external criterion variables (e.g., future student earnings) by removing mechanical correlations that arise from using the same students for both the VAM and the external criterion. Indeed, in some cases, the procedure may be counter-productive, because "if the goal is simply to document patterns in value added across ... teachers, jackknifing unnecessarily removes information ... [and] may even remove precisely the information that is most useful" (Bacher-Hicks & Koedel, 2023, p. 104). Because the bias in reliability occurs under the ideal condition of random assignment to clusters (Appendix A), and all observational VAM strategies—no matter how sophisticated—attempt to approximate this ideal, we hypothesize that similar results would obtain under alternative formulations. Nonetheless, future research should explore the extent to which the issues identified here are consistent across alternative VAM formulations.

4.3 Limitations

While our arguments are strengthened by the convergent evidence from analytic, simulation, and wide-ranging empirical results, we note several key limitations of our study:

1. Much of the VAM literature in the United States examines data from state longitudinal testing systems, for which item responses are generally unavailable to secondary researchers in public repositories. As a result, our empirical data come from program evaluations, which may differ in important but unknown ways from state testing contexts. Limiting our analysis to datasets in the US with common academic outcome measures, moderate to long tests, and relatively large sample sizes, leads us to conjecture that similar results would obtain in other

contexts. Furthermore, the magnitudes of cluster-by-item interaction variance are consistently large across our empirical datasets that span a range of outcomes, settings, and age groups, However, we view the replication of our approach with state test data to be a promising extension, particularly exploring the extent to which changes in test items across years may be related to VA reliability estimates derived from other methods.

2. While we explored pretest measurement error in our simulations and found that the results were generally consistent, we did not explore models that correct for pretest measurement error (Lockwood & McCaffrey, 2014). In general, measurement error reduces the predictive power of the pretest score. The extent to which pretest measurement error affects the reliability of VA estimates depends on the extent to which the pretest scores explain variation at different levels of the model. That is, increased residual variation at the student level would reduce estimated reliability, but increased residual variation at the cluster level would increase estimated reliability. Thus, the effects of pretest measurement error on VA reliability are complex. From a causal identification perspective, pretest measurement error may yield biased VA estimates to the extent that the model fails to fully control for pre-existing differences in student proficiency (Lockwood & McCaffrey, 2014). On the other hand, if students select into clusters based on *observed* pretest scores (such as exam schools with observed score cutoffs for admission), pretest measurement error adjustments may be counterproductive. While several of our empirical datasets contain item-level pretest data, software constraints in R limit us from estimating multilevel models with *both* latent pretests and item-level outcomes. That is, novel packages such as `galamm` (Sørensen, 2024) allow for cross-classified random effects models such as those explored in this study with either latent outcomes or latent covariates, but not both (Ø. Sørensen, personal communication, December 8, 2024), though fully Bayesian approaches, errors-in-variables corrections, or structural equation modeling approaches may provide alternatives (Bürkner, 2017; Lockwood & McCaffrey, 2020).

3. Our reliability estimates treat the variance components as known, when they are estimated with uncertainty (see Appendix D). Thus, when the number of clusters and/or items is low, estimates of σ_{ν}^2 may be unstable and point estimates of reliability may fail to capture the uncertainty of estimation, particularly in decision-making contexts. While beyond the scope of the present study, several researchers have proposed Bayesian or bootstrapping approaches that more easily allow for uncertainty estimates for variance components and may therefore provide an attractive approach in contexts with limited data (De Maeyer, 2021; Jiang & Skorupski, 2018; Jiang et al., 2022; LoPilato et al., 2015).

4. Our models assume that each item is equally discriminating with respect to the unobserved true score. This assumption is standard in Generalizability Theory applications but can be relaxed in factor analytic or IRT-based approaches that allow for a unique factor loading or item discrimination for each assessment item (McNeish & Wolf, 2020). Varying item discriminations could theoretically inflate σ_{ν}^2 because a constant improvement to latent academic achievement would manifest as differential improvements to item performance, even when the general non-linear scaling implied by a logit model is taken into account (Gilbert, Himmelsbach, et al., 2025, Appendix A). More flexible models that allow for varying item discriminations exist, but are computationally demanding (generally requiring MCMC estimation) and can be difficult to interpret (Bürkner, 2021; Gilbert, 2024; Gilbert, Young, et al., 2025; Gilbert, Zhang, et al., 2025; Petscher et al., 2020).

5. Our models assume a unidimensional construct, which is why we estimate separate models for the different outcomes from the subset of studies that contain more than one outcome measure. Multidimensional extensions of Generalizability Theory are possible and would allow for simultaneous consideration of multiple outcomes in a single model but are beyond the scope of the present study (Durvasula et al., 2006; Jiang & Skorupski, 2018; Vispoel et al., 2023).

6. Our arguments apply to estimates of VA reliability derived from Generalizability Theory formulas, whether based on scaled scores or item responses. When correlations between VA estimates are calculated directly and items differ between replications, the estimated correlation will inherently capture any cluster-by-item interaction variance. More concretely, when researchers calculate VA estimates for teachers across two separate years and correlate them, both the students and items typically vary across years and therefore the correlation appropriately captures both sources of variation (in addition to variation over time, which will likely deflate the estimated reliability further). Thus, our arguments about the importance of cluster-by-item interactions are most relevant to contexts in which estimating VA reliability is based on Generalizability Theory formulas rather than correlations between replications with varying items.

4.4 Conclusion

Identification of effective teachers and schools remains a promising and active area in educational research and practice. By adjusting for observed differences between students, VAMs provide useful estimates of putatively causal effects of teachers and schools when selection-on-observables assumptions are realistic. However, by failing to account for cluster-by-item interactions, standard approaches to VAM generally yield systematically inflated estimates of both the degree of variation between teachers or schools and the reliability of the VA estimates themselves. As a result, teacher and school effectiveness may be both less stable and more predictive of student outcomes than current evidence suggests. Thus, our understanding of the impact of teachers and schools on student learning will remain incomplete unless all sources of variation such as cluster-by-item interactions are appropriately accounted for.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135. https://doi.org/10.1086/508733

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Benjamin, W. (2024). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2024.2361337

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, *44*(8), 448–452. https://doi.org/10.3102/0013189X15618385

Amrein-Beardsley, A. (2014). *Rethinking Value-Added Models in Education* (1st ed.). Routledge. https://doi.org/10.4324/9780203409909

Amrein-Beardsley, A., Pivovarova, M., & Geiger, T. J. (2016). Value-added models: What the experts say. *Phi Delta Kappan*, *98*(2), 35–40. https://doi.org/10.1177/0031721716671904

Angrist, J., Hull, P., Pathak, P. A., & Walters, C. (2024). Credible school value-added with undersubscribed school lotteries. *Review of Economics and Statistics*, *106*(1), 1–19. https://doi.org/10.1162/rest_a_01149

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, *132*(2), 871–919. https://doi.org/10.1093/qje/qjx001

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, *24*(2), 443–483. https://doi.org/https://doi.org/10.1177/1094428119877457

Aslantas, I. (2020). Impact of contextual predictors on value-added teacher effectiveness estimates. *Education Sciences*, *10*(12), 390. https://doi.org/10.3390/educsci10120390

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis [Publisher: Wiley]. *Statistics in Medicine*, *36*(20), 3257–3277. https://doi.org/10.1002/sim.7336

Bacher-Hicks, A., & Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. In *Handbook of the Economics of Education* (pp. 93–134, Vol. 6). Elsevier. https://doi.org/10.1016/bs.hesedu.2022.11.002

Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. *Early Childhood Education Journal*, *51*(4), 717–732.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Béland, S., & Falk, C. F. (2022). A comparison of modern and popular approaches to calculating reliability for dichotomously scored items. *Applied Psychological Measurement*, *46*(4), 321–337. https://doi.org/10.1177/01466216221084210

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*, 1051–1074.

Bitler, M., Corcoran, S. P., Domina, T., & Penner, E. K. (2021). Teacher effects on student achievement and height: A cautionary tale. *Journal of Research on Educational Effectiveness*, *14*(4), 900–924. https://doi.org/10.1080/19345747.2021.1917025

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology*, *44*, 39–54.

Brehm, M., Imberman, S. A., & Lovenheim, M. F. (2017). Achievement effects of individual performance incentives in a teacher merit pay tournament. *Labour Economics*, *44*, 133–150. https://doi.org/10.1016/j.labeco.2016.12.008

Brennan, R. (2001). *Generalizability Theory*. Springer.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36*(5), 616–637. https://doi.org/10.3102/1076998610396887

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cabell, S. Q., Kim, J. S., White, T. G., Gale, C. J., Edwards, A. A., Hwang, H., Petscher, Y., & Raines, R. M. (2025). Impact of a content-rich literacy curriculum on kindergarteners' vocabulary, listening comprehension, and content knowledge. *Journal of Educational Psychology*, *117*(2), 153–175. https://doi.org/10.1037/edu0000916

Cárdenas-Hurtado, C. A., Moustaki, I., Chen, Y., & Marra, G. (2025). Generalized latent variable models for location, scale, and shape parameters. *Psychometrika*, 1–25. https://doi.org/10.1017/psy.2025.7

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics*, *40*(1), 35–68.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, *39*(5), 333–367. https://doi.org/10.3102/1076998614547576

Cawley, J., Heckman, J., & Vytlacil, E. (1999). On policies to reward the value added by educators. *Review of Economics and Statistics*, *81*(4), 720–727. https://doi.org/10.1162/003465399558436

Chan, W., & Hedges, L. V. (2022). Pooling interactions into error terms in multisite experiments. *Journal of Educational and Behavioral Statistics*, *47*(6), 639–665. https://doi.org/10.3102/10769986221104800

Chetty, R., Friedman, J., & Rockoff, J. (2014). Discussion of the American Statistical Association's statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy*, *1*(1), 111–113. https://doi.org/10.1080/2330443X.2014.955227

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-level assessments and teacher evaluation systems after the passage of the every student succeeds act: Some steps in the right direction* (tech. rep.). https://eric.ed.gov/?id=ED591993

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Cohodes, S. R. (2016). Teaching to the student: Charter school effectiveness in spite of perverse incentives. *Education Finance and Policy*, *11*(1), 1–42. https://doi.org/10.1162/EDFP_a_00175

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2023). Assessing licensure test performance and predictive validity for different teacher subgroups. *American Educational Research Journal*, *60*(6), 1095–1138. https://doi.org/10.3102/00028312231192365

Davenport, J. L., Kao, Y. S., Johannes, K. N., Hornburg, C. B., & McNeil, N. M. (2023). Improving children's understanding of mathematical equivalence: An efficacy study. *Journal of Research on Educational Effectiveness*, *16*(4), 615–642.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

De Maeyer, S. (2021). Generalizability theory with a Bayesian flavour. https://svendemaeyer.netlify.app/posts/2021-04-Generalizability/

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT: Incentives, selection, and teacher performance. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L. A., Gilbert, J., Kanopka, K., Kapoor, R., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2024). Solving the problem of data in psychometrics: An introduction to the Item Response Warehouse (IRW). https://doi.org/10.31234/osf.io/7bd54

Domingue, B. W., Kanopka, K., Trejo, S., Rhemtulla, M., & Tucker-Drob, E. M. (2022). Ubiquitous bias and false discovery due to model misspecification in analysis of statistical interactions: The role of the outcome's distribution and metric properties. *Psychological Methods*. https://doi.org/https://doi.org/10.1037/met0000532

Donnellan, E., Usami, S., & Murayama, K. (2023). Random item slope regression: An alternative measurement model that accounts for both similarities and differences in association with individual items. *Psychological Methods*. https://doi.org/10.1037/met0000587

Durvasula, S., Netemeyer, R. G., Andrews, J. C., & Lysonski, S. (2006). Examining the cross-national applicability of multi-item, multi-dimensional measures using generalizability theory. *Journal of International Business Studies*, *37*(4), 469–483. https://doi.org/10.1057/palgrave.jibs.8400210

Everson, K. C. (2017). Value-added modeling and educational accountability: Are we answering the real questions? *Review of Educational Research*, *87*(1), 35–70. https://doi.org/10.3102/0034654316637199

Gilbert, J. B. (2024). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. *Behavior Research Methods*, *56*(5), 5055–5067. https://doi.org/10.3758/s13428-023-02245-8

Gilbert, J. B. (2025). How measurement affects causal inference: Attenuation bias is (usually) more important than outcome scoring weights. *Methodology*, *21*(2), 91–122. https://doi.org/10.5964/meth.15773

Gilbert, J. B., Domingue, B. W., & Kim, J. S. (2025). Estimating causal effects on psychological networks using item response theory. *Psychological Methods*. https://doi.org/10.1037/met0000764

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2025). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*, *50*(1), 72–101. https://doi.org/10.3102/10769986241240085

Gilbert, J. B., & Soland, J. (2024). Mechanisms of effect size differences between researcher developed and independently developed outcomes: A meta-analysis of item-level data. https://doi.org/10.26300/8AXS-Y713

Gilbert, J. B., Young, W. S., Himmelsbach, Z., Ulitzsch, E., & Domingue, B. W. (2025). Conditional dependencies between response time and item discrimination: An item-level meta-analysis. https://doi.org/10.31234/osf.io/rp34w_v1

Gilbert, J. B., Zhang, L., Ulitzsch, E., & Domingue, B. W. (2025). Polytomous explanatory item response models for item discrimination: Assessing negative-framing effects in social-emotional learning surveys. *Behavior Research Methods*, *57*(4), 1–21. https://doi.org/10.3758/s13428-025-02625-2

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, *35*(2), 220–236. https://doi.org/10.3102/0162373712466938

Halpin, P., & Gilbert, J. (2024). Testing whether reported treatment effects are unduly dependent on the specific outcome measure used. https://doi.org/10.48550/ARXIV.2409.03502

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479. https://doi.org/10.1016/j.econedurev.2010.12.006

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*(4), 319–350. https://doi.org/10.1162/edfp.2009.4.4.319

Hawley, L. R., Bovaird, J. A., & Wu, C. (2017). Stability of teacher value-added rankings across measurement model and scaling conditions. *Applied Measurement in Education*, *30*(3), 196–212. https://doi.org/10.1080/08957347.2017.1316273

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Ho, A. D. (2024). Measurement must be qualitative, then quantitative, then qualitative again. *Educational Measurement: Issues and Practice*, *43*(4), 137–145. https://doi.org/10.1111/emip.12662

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In *Educational Measurement* (4th ed., pp. 187–220).

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. https://doi.org/10.1086/699018

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, *2*(4), 491–508. https://doi.org/10.1257/aeri.20200029

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *40*(2), 267–284. https://doi.org/10.3102/0162373718759600

Jeon, M.-J., Lee, G., Hwang, J.-W., & Kang, S.-J. (2009). Estimating reliability of school-level scores using multilevel and generalizability theory models. *Asia Pacific Education Review*, *10*(2), 149–158. https://doi.org/10.1007/s12564-009-9014-3

Jiang, Z., Raymond, M., DiStefano, C., Shi, D., Liu, R., & Sun, J. (2022). A monte carlo study of confidence interval methods for generalizability coefficient. *Educational and Psychological Measurement*, *82*(4), 705–718. https://doi.org/10.1177/00131644211033899

Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*, *50*(6), 2193–2214. https://doi.org/10.3758/s13428-017-0986-3

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* (tech. rep. No. ED540959). ERIC. https://eric.ed.gov/?id=ED540959

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3–26.

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2023). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*, *115*(1), 73–98.

Kim, J. S., Gilbert, J. B., Relyea, J. E., Rich, P., Scherer, E., Burkhauser, M. A., & Tvedt, J. N. (2024). Time to transfer: Long-term effects of a sustained and spiraled content literacy intervention in the elementary grades. *Developmental Psychology*, *60*(7), 1279–1297.

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Publications.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, *5*(1), 54–81. https://doi.org/10.1162/edfp.2009.5.1.5104

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record: The Voice of Scholarship in Education*, *116*(1), 1–21. https://doi.org/10.1177/016146811411600109

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Teachers College Record*, *107*(14), 99–118.

Koretz, D. (2008). *Measuring up*. Harvard University Press.

Leckie, G., Parker, R., Goldstein, H., & Tilling, K. (2024). Mixed-effects location scale models for joint modeling school value-added effects on the mean and variance of student achievement. *Journal of Educational and Behavioral Statistics*, *49*(6), 879–911. https://doi.org/10.3102/10769986231210808

Lee, Y. R., & Hong, S. (2019). The impact of omitting random interaction effects in cross-classified random effect modeling. *The Journal of Experimental Education*, *87*(4), 641–660. https://doi.org/10.1080/00220973.2018.1507985

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability*, *31*(3), 257–287. https://doi.org/10.1007/s11092-019-09303-w

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2023). How sensitive are the evaluations of a school's effectiveness to the selection of covariates in the applied value-added model? *Educational Assessment, Evaluation and Accountability*, *35*(1), 129–164. https://doi.org/10.1007/s11092-022-09386-y

Liu, J., & Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

Lockwood, J. R., & Castellano, K. E. (2015). Alternative statistical frameworks for student growth percentile estimation. *Statistics and Public Policy*, *2*(1), 1–9. https://doi.org/10.1080/2330443X.2014.962718

Lockwood, J. R., & McCaffrey, D. F. (2020). Recommendations about estimating errors-in-variables regression in Stata. *The Stata Journal: Promoting communications on statistics and Stata*, *20*(1), 116–130. https://doi.org/10.1177/1536867X20909692

Lockwood, J., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22–52.

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, *41*(2), 692–717. https://doi.org/10.1177/0149206314554215

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Manzi, J., San Martín, E., & Van Bellegem, S. (2014). School system evaluation by value added analysis under endogeneity. *Psychometrika*, *79*(1), 130–153. https://doi.org/10.1007/s11336-013-9338-0

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606. https://doi.org/10.1162/edfp.2009.4.4.572

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*, 2287–2305.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice*, *34*(4), 21–30. https://doi.org/10.1111/emip.12092

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82.

Morganstein, D., & Wasserstein, R. (2014). ASA statement on value-added models. *Statistics and Public Policy*, *1*(1), 108–110. https://doi.org/10.1080/2330443X.2014.956906

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: Journal of the Econometric Society*, 69–85.

Naumann, A., Hochweber, J., & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement*, *51*(4), 381–399.

Page, G. L., San Martín, E., Irribarra, D. T., & Van Bellegem, S. (2024). Temporally dynamic, cohort-varying value-added models. *Psychometrika*, *89*(3), 1074–1103. https://doi.org/10.1007/s11336-024-09979-0

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359–388.

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. (2020). Past perspectives and new opportunities for the explanatory item response model. *Annals of Dyslexia*, *70*, 160–179. https://doi.org/10.1007/s11881-020-00204-y

Pivovarova, M., Amrein-Beardsley, A., & Broatch, J. (2016). Value-added models (VAMs): Caveat emptor. *Statistics and Public Policy*, *3*(1), 1–9. https://doi.org/10.1080/2330443X.2016.1164641

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14.

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement*, *44*(1), 69–87. https://doi.org/10.1111/j.1745-3984.2007.00027.x

Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using Stata*. STATA Press.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17. https://doi.org/10.2307/2112482

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129. https://doi.org/10.3102/10769986029001121

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335. https://doi.org/10.3102/10769986020004307

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519. https://doi.org/10.1162/edfp.2009.4.4.492

Revelle, W., & Condon, D. M. (2019). Reliability from alpha to omega: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.589965

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571. https://doi.org/10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214. https://doi.org/10.1162/qjec.2010.125.1.175

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116. https://doi.org/10.3102/10769986029001103

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171. https://doi.org/10.3102/1076998611432174

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components* (2nd ed.). Wiley. https://doi.org/10.1002/9780470316856

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 1–15. https://doi.org/10.1348/000711008X398968

Sørensen, Ø. (2024). Multilevel semiparametric latent variable modeling in R with “galamm”. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2024.2385336

Taylor, E. S. (2023). Teacher evaluation and training. In *Handbook of the economics of education* (pp. 61–141, Vol. 7). Elsevier. https://doi.org/10.1016/bs.hesedu.2023.03.002

Thai, K.-P., Bang, H. J., & Li, L. (2022). Accelerating early math learning with research-based personalized learning games: A cluster randomized controlled trial. *Journal of Research on Educational Effectiveness*, *15*(1), 28–51. https://doi.org/10.1080/19345747.2021.1969710

Vispoel, W. P., Lee, H., Hong, H., & Chen, T. (2023). Applying multivariate generalizability theory to psychological assessments. *Psychological Methods*. https://doi.org/10.1037/met0000606

Wells, C. S., & Sireci, S. G. (2020). Evaluating random and systematic error in student growth percentiles. *Applied Measurement in Education*, *33*(4), 349–361. https://doi.org/10.1080/08957347.2020.1789139

Wiedermann, W., Zhang, B., Reinke, W., Herman, K. C., & Von Eye, A. (2024). Distributional causal effects: Beyond an “averagarian” view of intervention effects. *Psychological Methods*, *29*(6), 1046–1061. https://doi.org/10.1037/met0000533

Wulff, S. S. (2008). The equality of REML and ANOVA estimators of variance components in unbalanced normal classification models. *Statistics & Probability Letters*, *78*(4), 405–411. https://doi.org/10.1016/j.spl.2007.07.013

Ye, F., & Daniel, L. (2017). The impact of inappropriate modeling of cross-classified data structures on random-slope models. *Journal of Modern Applied Statistical Methods*, *16*(2), 458–484. https://doi.org/10.22237/jmasm/1509495900

Appendices

A Demonstration that Estimated VA Reliability is Upwardly Biased when Cluster-by-Item Interactions are Present but Omitted from the Model

Consider the following data-generating model that includes cluster-by-item interactions ν_{ik}:

$$y_{ijk} = u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk} \tag{21}$$

$$u_k \sim N(0, \sigma_u^2) \tag{22}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{23}$$

$$b_i \sim N(0, \sigma_b^2) \tag{24}$$

$$\nu_{ik} \sim N(0, \sigma_\nu^2) \tag{25}$$

$$e_{ijk} \sim N(0, \sigma_e^2), \tag{26}$$

where all random effects are assumed mutually independent. For clarity of exposition, we omit the grand intercept β_0 and the pretest covariate β_1 and suppose the data are balanced with $i = 1, ..., I$ items, $j = 1, ..., J$ students per cluster, and $k = 1, ..., K$ clusters. Thus, there are IJK total observations.

We are interested in the estimated reliability of u_k when, instead of Equation 21, we fit the following misspecified model that omits the cluster-by-item interactions ν_{ik}:

$$y_{ijk} = u_k + \theta_{jk} + b_i + e_{ijk} \tag{27}$$

$$u_k \sim N(0, \sigma_u^2) \tag{28}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{29}$$

$$b_i \sim N(0, \sigma_b^2) \tag{30}$$

$$e_{ijk} \sim N(0, \sigma_e^2). \tag{31}$$

We first establish ANOVA estimators for the variance components in Equation 27. In balanced designs, the closed-form ANOVA estimators of variance components are equivalent to REML estimation (Wulff, 2008). We define the following sums of squares:

$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{.jk} - \bar{y}_{i..} + \bar{y}_{...})^2 \tag{32}$$

$$SS_i = JK\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2 \tag{33}$$

$$SS_{j|k} = I\sum_{k=1}^{K}\sum_{j=1}^{J}(\bar{y}_{.jk} - \bar{y}_{..k})^2 \tag{34}$$

$$SS_k = IJ\sum_{k=1}^{K}(\bar{y}_{..k} - \bar{y}_{...})^2. \tag{35}$$

Using these, we define the mean squares:

$$MS_e = \frac{SS_e}{IJK} \tag{36}$$

$$MS_i = \frac{SS_i}{I-1} \tag{37}$$

$$MS_{j|k} = \frac{SS_{j|k}}{K(J-1)} \tag{38}$$

$$MS_k = \frac{SS_k}{K-1}. \tag{39}$$

The ANOVA estimators are defined in terms of these mean squares (Searle et al., 2006):

$$\widehat{\sigma}_e^2 = MS_e \tag{40}$$

$$\widehat{\sigma}_\theta^2 = \frac{MS_{j|k} - MS_e}{I} \tag{41}$$

$$\widehat{\sigma}_u^2 = \frac{MS_k - MS_{j|k}}{IJ}. \tag{42}$$

We ignore $\widehat{\sigma}_b^2$ here as it does not factor into the estimated relative reliability.

To understand the effects of model misspecification on estimated reliability, we next derive the expectations of these estimators when there is unmodeled cluster-by-item variance. That is, we find the expectations of the mean squares under Equation 27 when Equation 21 is the data-generating process.

We begin with $\mathbb{E}[MS_k]$. To derive this, it is convenient to develop alternate expressions of its constituent parts, $y_{..k}$ and $y_{...}$.

$$\overline{y}_{..k} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijk} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\big(b_i + \theta_{jk} + u_k + \nu_{ik} + e_{ijk}\big). \tag{43}$$

Then we regroup this expression as sums of each parameter.

$$\overline{y}_{..k} = \big(\tfrac{1}{I}\sum_{i=1}^{I} b_i\big) + \big(\tfrac{1}{J}\sum_{j=1}^{J}\theta_{jk}\big) + u_k + \big(\tfrac{1}{I}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ijk}\big). \tag{44}$$

Similarly for the overall (grand) sample mean:

$$\begin{aligned}
\overline{y}_{...} &= \frac{1}{IJK}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijk} && (45)\\
&= \frac{1}{IJK}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}\big(b_i + \theta_{jk} + u_k + \nu_{ik} + e_{ijk}\big), && (46)\\
&= \big(\tfrac{1}{I}\sum_{i=1}^{I} b_i\big) + \big(\tfrac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J}\theta_{jk}\big) + && (47)\\
&\big(\tfrac{1}{K}\sum_{k=1}^{K} u_k\big) + \big(\tfrac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJK}\sum_{k,i,j} e_{ijk}\big). && (48)
\end{aligned}$$

Recall that SS_k involves the squared difference of these two sample means. We now take this difference, grouping like terms:

$$\begin{aligned}
\big(\overline{y}_{..k} - \overline{y}_{...}\big) = \big(\tfrac{1}{J}\sum_{j=1}^{J}\theta_{jk} - \tfrac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J}\theta_{jk}\big) + \big(u_k - \tfrac{1}{K}\sum_{k=1}^{K} u_k\big) + && (49)\\
\big(\tfrac{1}{I}\sum_{i=1}^{I}\nu_{ik} - \tfrac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJ}\sum_{i,j} e_{ijk} - \tfrac{1}{IJK}\sum_{k,i,j} e_{ijk}\big). && (50)
\end{aligned}$$

This is equivalent to

$$\overline{y}_{..k} - \overline{y}_{...} = \big(\overline{\theta}_{.k} - \overline{\theta}_{..}\big) + \big(u_k - \overline{u}\big) + \Big(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\Big) + \big(\overline{e}_{..k} - \overline{e}_{...}\big). \tag{51}$$

Now recall that

$$SS_k = IJ\sum_{k=1}^{K}\Big[\big(\overline{y}_{..k} - \overline{y}_{...}\big)\Big]^2, \quad MS_k = \frac{SS_k}{K-1}, \tag{52}$$

and notice that

$$(\overline{y}_{..k} - \overline{y}_{...})^2 = (\overline{\theta}_{.k} - \overline{\theta}_{..})^2 + (u_k - \overline{u})^2 + \Big(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\Big)^2 + \tag{53}$$

$$\big(\overline{e}_{..k} - \overline{e}_{...}\big)^2 + \text{(cross terms)}, \tag{54}$$

where the cross-terms are all independent, so their expectations are zero. For example, $\mathbb{E}[(u_k - \overline{u})(\overline{e}_{..k} - \overline{e}_{...})] = \mathbb{E}[(u_k - \overline{u})]\,\mathbb{E}[(\overline{e}_{..k} - \overline{e}_{...})] = 0$. Therefore, we only need to consider the expectations of the squared terms. By linearity, we can consider the expectation of each squared difference separately. We then have

$$\mathbb{E}[(\overline{\theta}_{.k} - \overline{\theta}_{..})^2] = \frac{(K-1)\sigma_\theta^2}{J} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}[(\overline{\theta}_{.k} - \overline{\theta}_{..})^2] = I\sigma_\theta^2 \tag{55}$$

$$\mathbb{E}[(u_k - \overline{u})^2] = (K-1)\sigma_u^2 \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}[(u_k - \overline{u})^2] = IJ\sigma_u^2 \tag{56}$$

$$\mathbb{E}\Big[\Big(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\Big)^2\Big] = \frac{(K-1)\sigma_\nu^2}{I} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}\Big[\Big(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\Big)^2\Big] = J\sigma_\nu^2 \tag{57}$$

$$\mathbb{E}\Big[(\overline{e}_{..k} - \overline{e}_{...})^2\Big] = \frac{(K-1)\sigma_e^2}{IJ} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}\Big[(\overline{e}_{..k} - \overline{e}_{...})^2\Big] = \sigma_e^2, \tag{58}$$

where the first equation of each line is given by the standard ANOVA result for the squared difference between a group mean and a grand mean (Searle et al., 2006, Chapter 4). Now, by linearity, we have that

$$\mathbb{E}[MS_k] \ = \ IJ\sigma_u^2 \ + \ I\sigma_\theta^2 \ + \ J\sigma_\nu^2 \ + \ \sigma_e^2. \tag{59}$$

The expectations of the other mean squares can be derived similarly.

$$\mathbb{E}[MS_{j|k}] = I\sigma_\theta^2 + \sigma_e^2 \tag{60}$$

$$\mathbb{E}[MS_k] = IJ\sigma_u^2 \ + \ I\sigma_\theta^2 \ + \ J\sigma_\nu^2 + \sigma_e^2 \tag{61}$$

$$\mathbb{E}[MS_i] = JK\sigma_b^2 \ + \ J\sigma_\nu^2 \ + \ \sigma_e^2 \tag{62}$$

$$\mathbb{E}[MS_e] = \sigma_e^2 \ + \ \alpha\sigma_\nu^2 \tag{63}$$

where $\alpha = \frac{J(K-1)}{JK-1}$.

Given these expectations, the expectations of our variance estimators under the true model is then

$$\mathbb{E}[\widehat{\sigma}_e^2] = \mathbb{E}[MS_e] = \sigma_e^2 + \alpha\sigma_\nu^2 \tag{64}$$

$$\mathbb{E}[\widehat{\sigma}_\theta^2] = \mathbb{E}\left[\frac{MS_{j|k} - MS_e}{I}\right] = \sigma_\theta^2 - \frac{\alpha\sigma_\nu^2}{I} \tag{65}$$

$$\mathbb{E}[\widehat{\sigma}_u^2] = \mathbb{E}[\frac{MS_k - MS_{j|k}}{IJ}] = \sigma_u^2 + \frac{\sigma_\nu^2}{I}. \tag{66}$$

Our principal interest is in the estimated reliability of the cluster effect u_k:

$$\widehat{\rho} = \frac{\widehat{\sigma_u^2}}{\widehat{\sigma_u^2} + \frac{\widehat{\sigma_\theta^2}}{J} + \frac{\widehat{\sigma_e^2}}{IJ}} \tag{67}$$

By the delta method, the approximate expectation for this estimator is

$$\mathbb{E}[\widehat{\rho}] \approx \frac{\mathbb{E}[\widehat{\sigma_u^2}]}{\mathbb{E}[\widehat{\sigma_u^2}] + \mathbb{E}[\widehat{\frac{\sigma_\theta^2}{J}}] + \mathbb{E}[\widehat{\frac{\sigma_e^2}{IJ}}]} \tag{68}$$

$$= \frac{\sigma_u^2 + \frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2 - \frac{\alpha\sigma_\nu^2}{I}}{J} + \frac{\sigma_e^2 + \alpha\sigma_\nu^2}{IJ}} \tag{69}$$

$$= \frac{\sigma_u^2 + \frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{70}$$

Compared to Equation 17, we have added $\frac{\sigma_\nu^2}{I}$ to the numerator while the denominator is unchanged. Thus, the bias in reliability will be approximately equal to:

$$\mathbb{E}[\widehat{\rho}] - \rho \approx \frac{\frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}} \tag{71}$$

and will only be 0 when $\sigma_\nu^2 = 0$ or $I \to \infty$.

B Demonstration that Estimated VA Reliability is Upwardly Biased in the Standard Mean Score Model

We next show that if we use mean scores rather than individual item responses, the estimated reliability of u_k is similarly upwardly biased. Consider again the individual item responses and their variance:

$$y_{ijk} = u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk} \tag{72}$$

$$V(y_{ijk}) = \sigma_u^2 + \sigma_\theta^2 + \sigma_b^2 + \sigma_\nu^2 + \sigma_e^2. \tag{73}$$

When all students respond to the same set of items, we can ignore σ_b^2 because differences in item easiness do not affect relative differences in performance.

To obtain the variance of the student average score, $\text{post}_{jk} = \frac{1}{I}\sum_{i=1}^{I} y_{ijk}$, we divide both σ_ν^2 and σ_e^2 by I because the scores are averaged across items:

$$V(\text{post}_{jk}) = \sigma_u^2 + \sigma_\theta^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_e^2}{I}. \tag{74}$$

When fitting a standard VAM of post_{jk} in a multilevel model of students nested within clusters, instead of the true variance components σ_u^2 and σ_θ^2 we instead estimate $\widehat{\sigma_u^2}$ and $\widehat{\sigma_\theta^2}$ where $\frac{\sigma_\nu^2}{I}$ and $\frac{\sigma_e^2}{I}$ are absorbed into the cluster and student components, respectively:

$$\mathbb{E}[\widehat{\sigma_u^2}] = \sigma_u^2 + \frac{\sigma_\nu^2}{I} \tag{75}$$

$$\mathbb{E}[\widehat{\sigma_\theta^2}] = \sigma_\theta^2 + \frac{\sigma_e^2}{I}. \tag{76}$$

Plugging these estimates into the VA reliability formula based on student average scores yields:

$$\begin{aligned}\widehat{\rho}_{\text{mean}} &= \frac{\widehat{\sigma_u^2}}{\widehat{\sigma_u^2} + \frac{\widehat{\sigma_\theta^2}}{J}} && (77)\\ &= \frac{\sigma_u^2 + \frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. && (78)\end{aligned}$$

Compared to Equation 17, we have increased the numerator by $\frac{\sigma_\nu^2}{I}$, yielding the same upwardly biased reliability observed in Appendix A for the analysis of the item-level data.

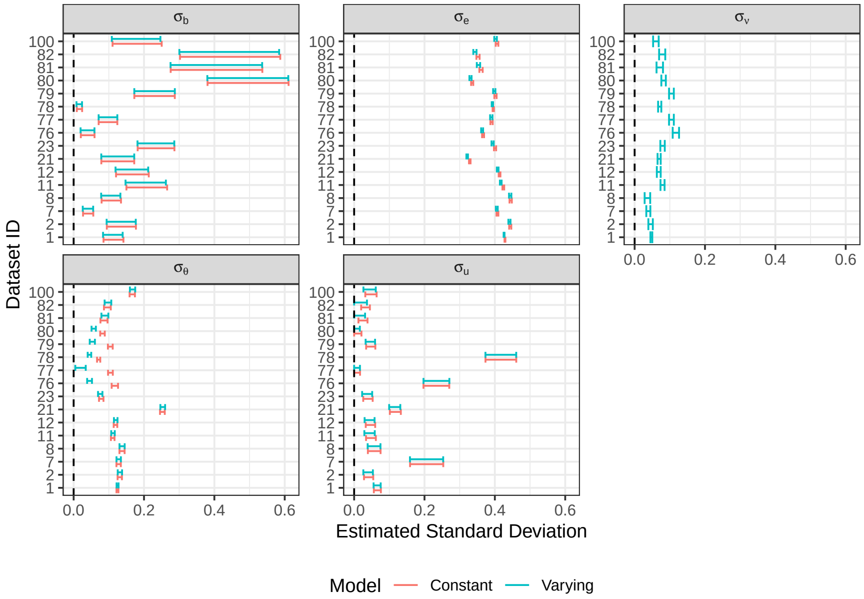
C R Code to Fit the Item Level Models

The R code below shows how to fit Equations 5 and 11 using `lme4`. We assume outcome variable `resp`, person identifier `id`, item identifier `item`, cluster identifier `cluster_id`, and baseline variable `pretest`.

```r
# item main effects
lmer(resp ~ pretest + (1|item) + (1|id) + (1|cluster_id), dat)

# item interaction effects
lmer(resp ~ pretest + (1|item) + (1|id) + (1|cluster_id) +
    (1|cluster_id:item), dat)

# sensitivity checks
# random slope for pretest
lmer(resp ~ pretest + (1|item) + (1|id) + (1|cluster_id) +
    (pretest|cluster_id:item), dat)

# logit model
# (assuming resp is 0/1)
glmer(resp ~ pretest + (1|item) + (1|id) + (1|cluster_id) +
    (1|cluster_id:item), dat, family = binomial)
```

D Additional Empirical Results

Figure D.1: Profile Confidence Intervals for Residual Standard Deviations in the Empirical Data



The x-axis shows the 95% profile CI and the y-axis shows the dataset ID. The panels are faceted by variance component and color-coded by model.

Table D.1: Likelihood Ratio Tests Comparing Constant and Varying VAMs

Dataset ID	Model	LL	AIC	BIC	χ^2
1	Constant	-134779.25	269572.50	269644.83	
1	Varying	-134375.12	268766.24	268848.90	808.26***
2	Constant	-27531.23	55076.45	55137.21	
2	Varying	-27478.29	54972.59	55042.03	105.86***
7	Constant	-26168.87	52351.73	52413.20	
7	Varying	-26130.96	52277.91	52348.15	75.82***
8	Constant	-23050.87	46115.74	46175.24	
8	Varying	-23035.05	46086.10	46154.11	31.63***
11	Constant	-34264.45	68542.90	68605.76	
11	Varying	-33725.53	67467.06	67538.91	1077.84***
12	Constant	-32676.58	65367.16	65429.97	
12	Varying	-32331.58	64679.17	64750.96	689.99***
21	Constant	-17892.26	35798.52	35859.56	
21	Varying	-17629.16	35274.32	35344.08	526.2***
23	Constant	-17499.34	35012.68	35071.65	
23	Varying	-17351.25	34718.51	34785.90	296.17***
76	Constant	-14517.28	29048.55	29107.46	
76	Varying	-14487.43	28990.86	29058.18	59.69***
77	Constant	-13997.10	28008.21	28065.88	
77	Varying	-13995.07	28006.15	28072.05	4.06*
78	Constant	-44990.41	89994.81	90060.60	
78	Varying	-44937.91	89891.82	89967.00	104.99***
79	Constant	-16229.80	32473.60	32531.91	
79	Varying	-16190.15	32396.30	32462.95	79.3***
80	Constant	-8592.45	17198.90	17255.65	
80	Varying	-8517.53	17051.07	17115.92	149.83***
81	Constant	-5291.74	10597.48	10649.46	
81	Varying	-5226.40	10468.80	10528.22	130.67***
82	Constant	-5176.98	10367.96	10420.11	
82	Varying	-5081.84	10179.68	10239.28	190.28***
100	Constant	-14390.47	28794.94	28851.94	
100	Varying	-14295.13	28606.26	28671.40	190.69***

Notes: LL = log-likelihood, $^{*}p < .05,^{**}p < .01,^{***}p < .001$. Because the null hypothesis is on the boundary of the parameter space (estimated variances cannot be negative), the reported p-value is divided by two (Rabe-Hesketh & Skrondal, 2022).

E Covariates Available in Empirical Datasets

Table E.1: Available Covariates in Empirical Datasets

Dataset ID	None	Gender	Race	SES	Age	Other
1	✓					
2		✓	✓	✓		✓
7		✓	✓	✓		✓
8		✓	✓	✓		✓
11		✓	✓	✓		✓
12		✓	✓	✓		✓
21	✓					
23	✓					
76		✓			✓	
77		✓				
78		✓				
79		✓				
80		✓				
81		✓				
82		✓				
100	✓					

Notes: We consider covariates other than baseline scores.