Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1173

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 a case study and 41 measures from 25 studies with item-level outcome data, we show how standard VAMs overstate reliability by an average of .12 on the 0-1 reliability scale (median = .09, SD = .13) and provide standard deviations of teacher/school effects that are on average 22% too large (median = 7%, SD = 41%). We discuss how imprecision due to heterogeneous VA effects across items attenuates effect sizes, obfuscates comparisons across studies, and causes instability over time. 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: April 2025

Suggested citation: Gilbert, Joshua B., Zachary Himmelsbach, Luke W. Miratrix, Andrew D. Ho, and Benjamin W. Domingue. (2025). Item-Level Heterogeneity in Value Added Models: Implications for Reliability, Cross-Study Comparability, and Effect Sizes. (EdWorkingPaper: 25 -1173). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ez4q-fs31

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

April 4, 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 a case study and 41 measures from 25 studies with item-level outcome data, we show how standard VAMs overstate reliability by an average of .12 on the 0-1 reliability scale (median = .09, SD = .13) and provide standard deviations of teacher/school effects that are on average 22% too large (median = 7%, SD = 41%). We discuss how imprecision due to heterogeneous VA effects across items attenuates effect sizes, obfuscates comparisons across studies, and causes instability over time. 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

JEL Codes: I21, C51, H75

Corresponding author: joshua_gilbert@g.harvard.edu

The authors wish to thank Mike Hardy, Doug Mosher, Eric Taylor, Alex Bolves, Lily An, and seminar participants at Harvard University 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 case study dataset from Brandt (2023) is available at the following URL: https://dataverse.harvard.edu/file.xhtml?fileId=5429719&version=1.0. The RCT datasets from Gilbert, Himmelsbach, et al. (2024) are available at the following URL: https://doi.org/10.7910/DVN/C4TJCA. The datasets are also available in the Item Response Warehouse (IRW, B. 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; D. D. 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 (C. K. Jackson, 2018; C. K. 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 25 years (Amrein-Beardsley et al., 2016; Association, 2015; 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; 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 (i.e., show low reliability), 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 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 both a case study of a large administrative dataset from secondary schools in Tanzania and 41 outcomes across 25 empirical studies in education with item-level outcome data and baseline variables. We find that cluster-by-item interactions are both prevalent and large in magnitude, leading to an average overestimation of VA reliability by .12 (SD = .12) on the 0-1 reliability scale and provide standard deviations of VA effects that are on average 20% too large (median = 7%, SD = 41%). Thus, researchers using standard approaches to VAM may be severely 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 (e.g., Castellano and 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, and others (see Koedel et al., 2015 for a review), we use the simple framework of Equation 1 throughout this study for clarity of exposition and because our arguments about the reliability of the cluster VA effect u_k do not depend on the specific VAM formulation used. 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 effect 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 (Bollen, 1989; 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 $\frac{\sigma_\theta^2}{J}$, where J is the number of students in cluster k. Thus, the observed variance of VA estimates for clusters of size J 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, 1992, 2001; Gleser et al., 1965). 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). Note that some research refers to the "stability" or "persistence" of VA effects rather than VA reliability. We use the term reliability in this study because it is more general and provides a framework for capturing the consistency of measurements across replications, however defined (e.g., over time, with different students, with different items, etc.).

Research on VA reliability shows mixed results, with some studies showing high estimates of VA reliability and others showing low estimates. For example, Briggs and Weeks (2011) show that school VA estimates have generally high reliability, with strongly correlated results across estimation methods and over time, whereas Yeh (2012) argues that the low reliability of teacher VA estimates is a serious problem for their use in teacher accountability systems due to high misclassification rates. Methodological considerations may partially explain these discrepant findings, as several studies demonstrate the sensitivity of VAMs to the choice of outcome measure, model specification, measurement error adjustments, and included covariates (Ehlert et al., 2014; D. D. Goldhaber et al., 2013; Levy et al., 2019; Lockwood et al., 2007; McCaffrey et al., 2009; Newton et al., 2010; Tekwe et al., 2004; Van De Grift, 2009), as well as idiosyncratic features such as the timing of a test within

a school year (Atteberry & Mangan, 2020; Papay, 2011). Closely related is the extent to which VA effects are stable from year to year. Again, the literature is mixed, with some studies showing strongly correlated teacher VA effects over several years (Kersting et al., 2013; Konstantopoulos & Chung, 2011), and others showing that less than half of VA effects persist over time (Andrabi et al., 2011; Kinsler, 2012). Similarly, some research has examined how teacher VA effects may vary across the students within a classroom (D. Goldhaber & Hansen, 2013; Lockwood & McCaffrey, 2009; Loeb et al., 2014).

Taken together, these findings suggest the importance of appropriately assessing VA reliability and averaging estimates over sufficient replications (e.g., averaging teacher VA estimates over several years) to obtain the desired level of precision. As we will show, consideration of item-specific VA effects may help to explain the variation in these results.

1.3 VAMs with Item-Level Data

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 and polytomous 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., 2024; 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; Holland, 1990), 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 maintain focus on single-occasion estimates of reliability.

The reliability of VA estimates has received much commentary in the education research literature, as described previously. However, 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, 1992, 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):

1. 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?

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

3. 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 of a case study of Tanzanian secondary schools and 41 additional datasets in education that contain both item responses and baseline scores.

2 Methods

2.1 Analytic Derivation and Simulation

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 verify these analytic results with a simulation study under a range of more complex conditions, including dichotomous item responses and pretest measurement error, in Appendix C.

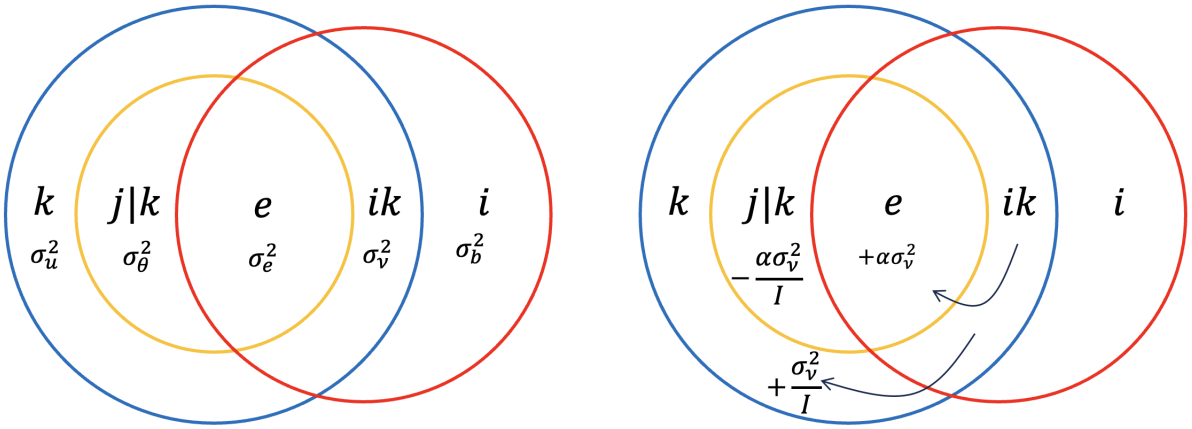
We can also understand these results on an intuitive level. Because the outcome is continuous, the total variance in y_{ijk} is constant across models and σ_ν^2 is distributed among the other variance components included in the model (Chan & Hedges, 2022; Lee & Hong, 2019; Shi et al., 2010; Ye & Daniel, 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 Empirical Application

2.2.1 Data Sources

We first analyze data from Brandt (2023), who examines a large administrative dataset from Tanzanian secondary schools. We use these data for our case study because of their large size and nation-wide scale. The outcome comprises secondary school test scores from the Form Two National Assessment (FTNA), a standardized assessment administered by the Tanzanian government, from about 650,000 students in about 5,000 secondary schools. A similar standardized test in primary school, the Primary School Leaving Examination (PSLE), serves as the primary control variable. We examine FTNA scores on 9 items (biology, chemistry, civics, English, geography, history, Swahili, math, and physics) and PSLE scores on 5 items (English, Swahili, math, science, and society).

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.

All items are coded on a 0-4 Likert scale representing a letter grade in each subject (F = 0, D = 1, C = 2, B = 3, A = 4). While the letter grade for each test subject is itself derived from yet more individual test items, the dataset only provides the subject grades, which we treat as our "items" for the purposes of this analysis. For simplicity, we limit our analysis to students with complete primary and secondary school test score data. We generate scaled scores from the FTNA and PSLE letter grades using a graded response model to serve as our outcome in the conventional VAM and our control variable in all models, respectively (see Shores and Student, 2024).

We then 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 (B. Domingue et al., 2024; Gilbert, Himmelsbach, et al., 2024). Here, we limit our analysis to the 41 outcomes from 25 studies in education in which the subjects are clustered in a higher-level unit such as classrooms or schools. Table 1 summarizes the datasets and shows a wide range of regions, outcomes, and age groups. When item-level data are available for the baseline measure, we construct scaled scores using a One-Parameter Logistic (1PL) IRT model for dichotomous responses and a Partial Credit

Model for polytomous responses. We examine only immediate post-intervention data in the eight 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.

Table 1: Empirical Datasets

Dataset	N	K	I	$\frac{N}{K}$	Cluster	Location	Age	Outcome
1: Gilbert et al. (2023)	7797	110	30	70.88	school	USA	G3	Reading Comprehension
2: Kim et al. (2023)	2174	30	20	72.47	school	USA	G2	Reading Comprehension
5: Woods-Townsend et al. (2021)	2486	37	7	67.19	school	UK	Adolescents	Health Literacy
6: Bruhn et al. (2016)	15395	842	10	18.28	school	Brazil	Adolescents	Financial Literacy
7: Kim et al. (2024)	1352	30	36	45.07	school	USA	G3	Vocabulary
8: Kim et al. (2024)	1303	30	29	43.43	school	USA	G3	Reading Comprehension
10: Kim et al. (2021)	4834	58	20	83.34	school	USA	G1-G2	Reading Self Concept
11: Kim et al. (2021)	2565	30	24	85.50	school	USA	G1	Vocabulary
12: Kim et al. (2021)	2580	30	24	86.00	school	USA	G2	Vocabulary
13: Romero et al. (2020)	3381	178	20	18.99	school	Liberia	Elementary	Literacy
14: Romero et al. (2020)	3381	178	44	18.99	school	Liberia	Elementary	Math
15: Romero et al. (2020)	3381	178	10	18.99	school	Liberia	Elementary	Raven's Progressive Matrices
16: de Barros et al. (2024)	3202	292	32	10.97	school	India	G4	Math
17: A. Duflo et al. (2024)	17344	498	21	34.83	school	Ghana	G1-G3	Math
18: A. Duflo et al. (2024)	17344	498	21	34.83	school	Ghana	G1-G3	English
19: A. Duflo et al. (2024)	17331	498	21	34.80	school	Ghana	G1-G3	Local Language
21: Davenport et al. (2023)	3671	172	13	21.34	class	USA	G5	Math
22: Berry et al. (2018)	5290	135	10	39.19	school	Ghana	Adolescents	Saving Attitudes
23: Bang et al. (2023)	886	41	38	21.61	class	USA	K-G1	Math
24: Llauradó et al. (2014)	495	20	13	24.75	school	Spain	Elementary	Dietary Behavior
25: Schreinemachers et al. (2020)	775	30	15	25.83	school	Nepal	8-12	Food Preferences
26: Schreinemachers et al. (2020)	775	30	15	25.83	school	Nepal	8-12	Food Knowledge
31: E. Duflo et al. (2015)	11893	400	6	29.73	school	India	Elementary	Academic Achievement
32: Maruyama (2022)	3619	232	20	15.60	school	El Salvador	G7	Math
35: Persson et al. (2020)	1152	59	12	19.53	class	Sweden	High School	Democratic Values
36: Persson et al. (2020)	1108	59	7	18.78	class	Sweden	High School	Political Knowledge
41: Mohohlwane et al. (2023)	3068	180	134	17.04	school	South Africa	Early Elementary	Oral Reading Fluency
46: Glatz et al. (2023)	120	9	42	13.33	class	Netherlands	G1	Language
47: Glatz et al. (2023)	123	10	44	12.30	class	Netherlands	G1	Math
56: Sebele et al. (2023)	2307	74	4	31.18	school	Liberia	Preschool	Literacy
64: Zhao et al. (2023)	4041	216	9	18.71	school	Jordan	Preschool	Social Emotional Learning

Continued on next page

Dataset	N	K	I	$\frac{N}{K}$	Cluster	Location	Age	Outcome
68: Banerjee et al. (2017)	5974	399	35	14.97	school	India	G1-G4	Hindi
69: Banerjee et al. (2017)	5966	399	30	14.95	school	India	G1-G4	Math
74: Gilbert, Kim, and Miratrix (2024)	1225	29	12	42.24	school	USA	G2	Vocabulary
76: Thai et al. (2022)	428	20	78	21.4	classroom	USA	K	Math
77: Cabell et al. (2025)	1075	47	26	22.9	school	USA	K	Language Fundamentals
78: Cabell et al. (2025)	1075	47	186	22.9	school	USA	K	Vocabulary
79: Cabell et al. (2025)	1100	47	30	23.4	school	USA	K	Narrative Language
80: Cabell et al. (2025)	1075	47	35	22.9	school	USA	K	Vocabulary
81: Cabell et al. (2025)	1075	47	18	22.9	school	USA	K	Science
82: Cabell et al. (2025)	1075	47	18	22.9	school	USA	K	Social Studies

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. (2024). We include the original dataset IDs in our tables and figures to facilitate replicability and comparability with the source study.

2.2.2 Empirical Models

We fit three models to each dataset: (1) traditional VAM with scaled scores, (2) item-level VAM assuming constant item effects, and (3) item-level VAM allowing for cluster-by-item interactions. We 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). Thus, the empirical models are specified in reduced form as follows, in which T_k is the treatment indicator and $\overline{\text{pre}}_k$ is the average pretest score for students within cluster k, with normal distributions on all random effects:

$$\text{Traditional VAM: } \text{post}_{jk} = \beta_0 + \beta_1 \text{pre}_{jk} + \gamma_1 T_k + \gamma_2 \overline{\text{pre}}_k + u_k + \theta_{jk} \tag{19}$$

$$\text{Constant Item VAM: } y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + \gamma_1 T_k + \gamma_2 \overline{\text{pre}}_k + u_k + \theta_{jk} + b_i + e_{ijk} \tag{20}$$

$$\text{Varying Item VAM: } y_{ijk} = \beta_0 + \beta_1 \text{pre}_{jk} + \gamma_1 T_k + \gamma_2 \overline{\text{pre}}_k + u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk}. \tag{21}$$

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. For our case study analysis of the data from Brandt (2023), the indicator for whether the school is public or private serves as an analog to the treatment indicator in the other datasets. We primarily focus on the differences between Equations 20 and 21.

3 Results

3.1 Tanzanian Secondary Schools

Our analytic sample includes 5,697,342 item responses from 633,038 students across 4,721 schools. The marginal internal consistencies (Dimitrov, 2003) of the FTNA and PSLE scaled scores are .92 and .80, respectively, and the student-level correlation between pre- and post-test scaled scores is $r = .49$. Adjusting this correlation for the reliability of each score yields a disattenuated correlation of $r = \frac{.49}{\sqrt{.92}\sqrt{.80}} = .57$. Year-to-year correlations for school-age children on standardized tests are typically stronger in United States data, around 0.75 to 0.85 for adjacent grades (e.g., Castellano and Ho, 2013; Pollack et al., 2005). Exploratory factor analysis of the FTNA and PSLE show strong evidence of unidimensionality, as the first factor explains the vast majority of the total variance. We include additional descriptive statistics and psychometric analyses for the FTNA and PSLE items in our supplement.

Table 2 shows the model results. We find relatively large cluster-by-item interaction variance, representing about one third of the total item-level VA variance. Substantively, the estimated SD of the cluster-by-item interactions of $\sqrt{.052} = .23$ means that 95% of the variation in school VA on individual items is within $\pm .46$ points (on the 0-4 scale) of the overall VA. Accordingly, the AIC, BIC, and log-likelihood fit statistics show that Model 3 provides a vastly better fit to the data than Model 2. In line with our simulation results (Appendix C), fixed effect coefficients and standard errors are not affected, and the variance components of Model 2 behave as we would expect from the analytic results in Appendix A. Namely, $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ increase whereas $\widehat{\sigma}_\theta^2$ decreases.

Figure 2 shows the implications of the cluster-by-item interactions for VA reliability, in which the x-axis shows school sample size and the y-axis shows estimated VA reliability derived from the variance components in Table 2, with the lines color-coded by model. Omitting the cluster-by-item interaction term from the model inflates estimated reliability, as expected, and this holds for both the item-level VAM (Model 2) and the traditional VAM using scaled scores as the outcome (Model 1). To illustrate the policy implications of Figure 2, consider the number of students needed to obtain a VA estimate with reliability .90, depicted as a horizontal dashed black line in the figure. Under the scaled score model, the policymaker will erroneously conclude that about 30 students are required, when in truth, the magnitude of cluster-by-item interactions means that about 60 students are required. We include additional discussion of the interpretations of the various reliability estimates in this dataset in Appendix D. Namely, Models 1 and 2 capture the expected correlation between VA estimates when students vary but items remain fixed; Model 3 captures the expected correlation between VA estimates when *both* students and items vary. As we argue in Section 1, generally, the latter estimate is more meaningful when the goal is to generalize VA effects to the broader content domain represented by the items.

3.2 Large-Scale Application

Our analytic sample of 41 outcomes from 25 studies contains 165,241 respondents (some of whom are included more than once because some studies include multiple outcome measures), 1,219 items, and 3,476,021 item responses. In general, outcome internal consistency is high (median = .82), item discrimination is high (median = 1.32), and item discriminations are somewhat variable within outcomes (median SD = .51).

Figure 3 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 the vast majority 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

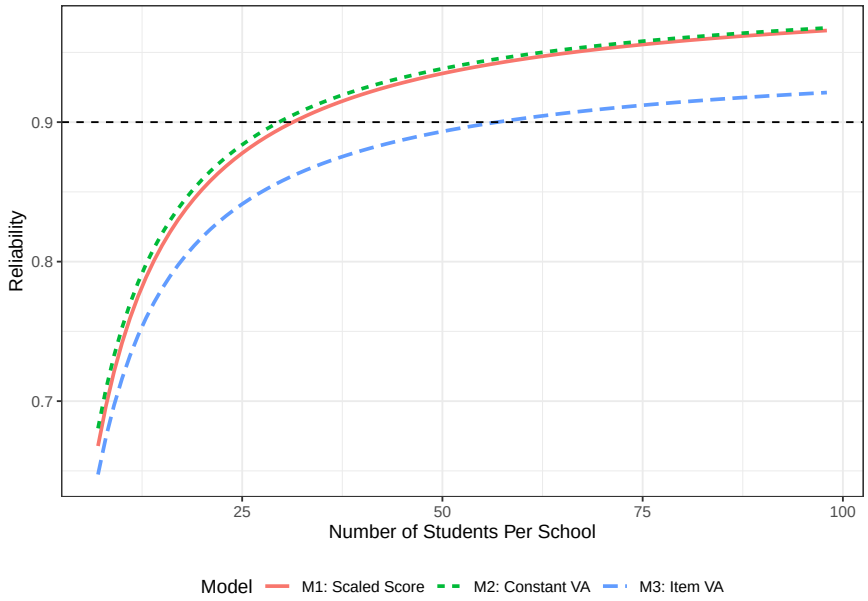
Table 2: VAMs Fit to Brandt (2023) Data

	M1: Scaled Scores	M2: Item Main Effects	M3: Item Interactions
Intercept	-0.069 $(0.007)^{***}$	1.031 (0.536)	1.031 $(0.157)^{***}$
Baseline	0.464 $(0.001)^{***}$	0.450 $(0.001)^{***}$	0.450 $(0.001)^{***}$
Mean Baseline	0.084 $(0.012)^{***}$	0.153 $(0.011)^{***}$	0.153 $(0.011)^{***}$
1 = Private	0.786 $(0.014)^{***}$	0.791 $(0.013)^{***}$	0.792 $(0.013)^{***}$
AIC	1376361.496	12487092.128	12021003.212
BIC	1376429.646	12487200.572	12021125.212
Log Likelihood	-688174.748	-6243538.064	-6010492.606
Num. obs.	633038	5697342	5697342
K	4721	4721	4721
σ_u^2	0.144	0.121	0.115
σ_e^2	0.502	0.411	0.367
N		633038	633038
I		9	9
σ_θ^2		0.353	0.358
σ_b^2		2.587	0.220
$K \times I$			42489
σ_ν^2			0.052

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Notes: Model 1 is based on a scaled score derived from a graded response model; Models 2 and 3 use the individual item responses. K = number of schools, N = number of students, I = number of items. The baseline score is the PSLE scaled score derived from a graded response model. Mean Baseline is the school-average pretest score.

Figure 2: Model-Implied Reliability of School VA Estimates from Brandt (2023) Data



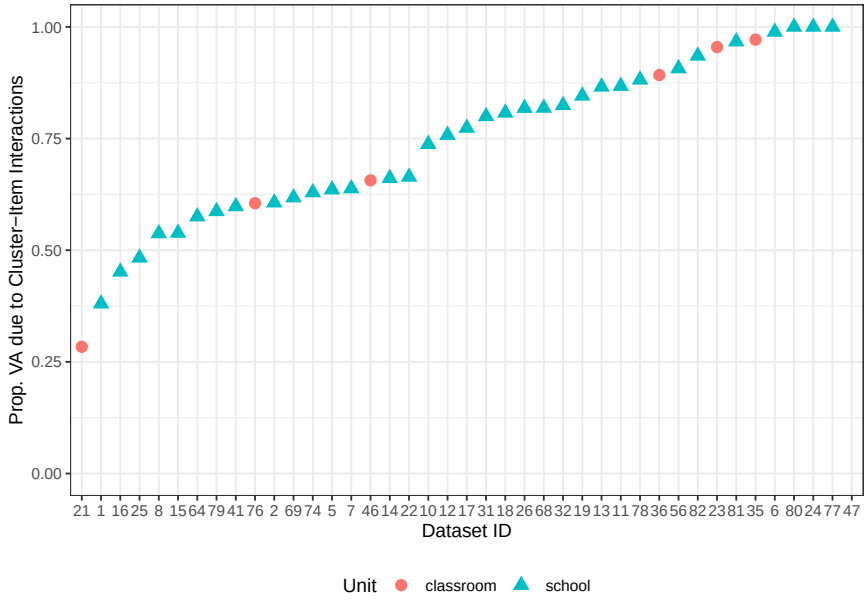
The y-axis shows the estimated reliability of a school VA estimate and the x-axis shows the number of students per school and ranges from the 10th to 90th percentile of observed school sample sizes. The lines are color coded by model. Model 1 is based on a scaled score derived from a graded response model. Models 2 and 3 use the individual item responses and assume constant VA and cluster-by-item interactions, respectively. The horizontal dashed line shows the number of students required to achieve a reliability of .90.

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., 2024; Halpin & Gilbert, 2024). We include tables of the full model results for each dataset in our supplement.

Figure 4 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 = .12, median = .09, SD = .13, on the 0-1 reliability scale). The differences in reliability are sometimes substantial, with many datasets showing inflation in estimated VA reliability of .20 or greater when cluster-by-item interactions are not accounted for. Accordingly, Figure 5 shows that the estimated SDs of the cluster

effects ($\widehat{\sigma}_u$) derived from the main effects models are on average 22% greater than those of the interaction models (median = 7%, SD = 41%).

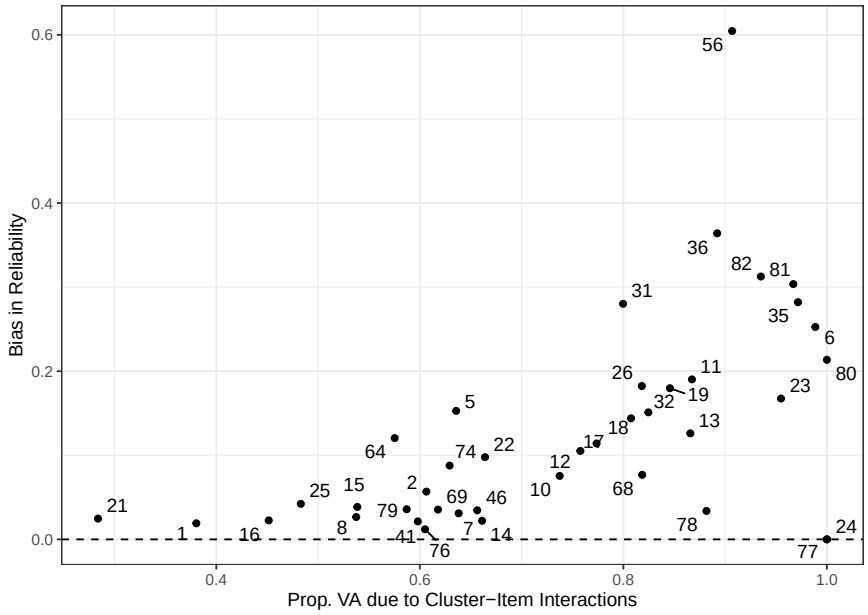
Figure 3: Proportion of Item-Level VA Variance Due to Cluster-by-Item Interactions in Empirical Datasets



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. Dataset 47 is missing because reliability from both models is estimated at 0.

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 only slightly inflates estimated reliability compared to the item-level model assuming constant VA effects (mean difference = .016, $p = .28$), whereas estimated reliability is significantly inflated compared to the cluster-item interaction model (mean difference = .138, $p < .001$). 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 4: Estimated Bias in VA Reliability in Empirical Datasets



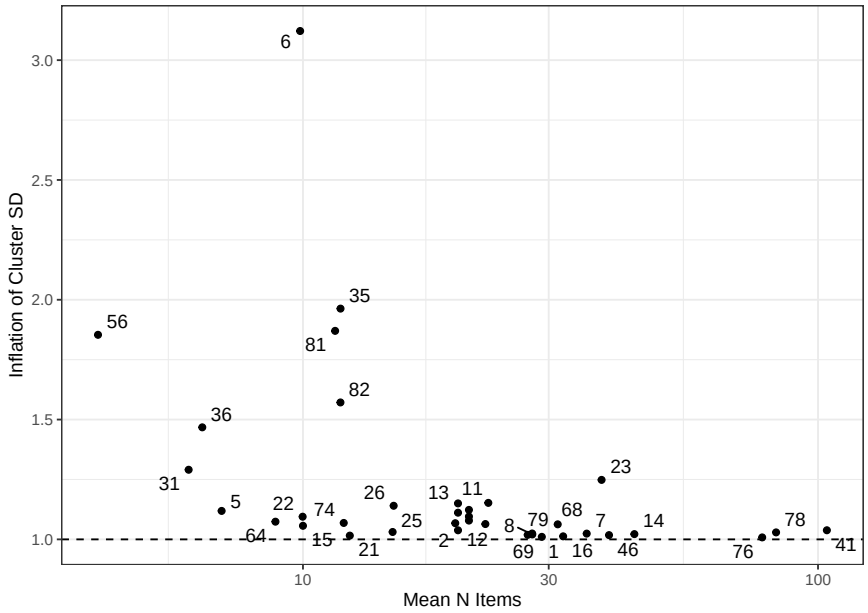
The x-axis shows the proportion of item-level VA variance due to cluster-by-item interactions ($\frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2}$) and the y-axis shows the estimated bias in reliability. The points are labeled by dataset ID. Dataset 47 is omitted because reliability from both models is estimated at 0.

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 two 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 and (2) non-linearity and ceiling/floor effects induced by the categorical item responses.

First, consider the assumption that the relationship β_1 between pre_{jk} and y_{ijk} is constant across all combinations of items and clusters (e.g., C. D. Jackson, 2013). 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 those specific competencies. This would occur to the extent that added instruction

Figure 5: Estimated Bias in VA SDs in Empirical Datasets



The x-axis shows the mean number of items answered by students and the y-axis shows the estimated bias 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. Three datasets are omitted because $\widehat{\sigma}_u$ from the constant effects models are estimated at almost exactly 0.

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}. This could potentially reduce σ^2_ν. Applying this more flexible model to our empirical data shows near identical results to our primary analyses: the difference in VA reliability estimates between the two methods is on average less than .01 and the mean proportion of item-level VA variance due to cluster-by-item interactions is 72%, compared to 75% for our main specification. Thus, differential prediction of item performance by pretest appears to be an unlikely explanation for our findings.

Second, all items from our empirical applications are categorical (dichotomous or polytomous) 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^2_ν 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 categorical item responses in two ways. First, we replicate our simulation study (Appendix C) with dichotomous items (and pretest measurement error) and find that the pattern of results is identical to those using the models with continuous items. Second, we fit cross-classified logit models to the 36 of our empirical datasets with binary item responses because logit models can correct for ceiling and floor effects and related scaling issues (B. W. 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 68% compared to 75% in the linear models. Thus, the categorical item responses are unlikely to confound the large estimates of σ^2_ν observed in the linear models.

4 Discussion

4.1 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 and their persistence over time, 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.

Our findings have several important 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 account for about 33% in a large administrative dataset from Tanzania (Table 2) and typically exceed 50% in our sample of 41 outcomes from 25 empirical studies in education (Figure 3).

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 (Figure 2). Thus, assessments longer than standard VAMs predict 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.

4. 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) (Appendix D).

5. 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, 2024a; Hedges, 1981). 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.

6. Because cluster-by-item interactions are not well conceptualized in VAMs for non-cognitive variables, our findings may partially explain why teachers' VA on behavioral measures appears more strongly correlated with students' long-term outcomes than teachers' test-score VA. For example, C. K. Jackson (2018) uses absences, suspensions, grades, and grade repetition as a proxy for non-cognitive skills in a VAM framework. Unlike state standardized test items, Jackson's non-cognitive "items" remain constant year over year and thus the effect of potential cluster-by-item interactions does not reduce the reliability of the VA estimates, potentially resulting in less attenuation of the effect compared to test score VA estimates.

7. 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 4 and 5, 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 (B. 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 $\frac{\sigma_\nu^2}{\sigma_\nu^2+\frac{\sigma_\theta^2}{J}+\frac{\sigma_\epsilon^2}{J}}$. For example, in the Brandt (2023) data, an individual cluster-by-item interaction has a reliability of .65 for a school with 25 students and .78 for a school with 50 students.

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 could provide important insights into the sources of VA effects in a broad range of real-world 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. 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).

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 mostly come from global program evaluations, which may differ in important but unknown ways from state testing contexts. While the consistently large magnitudes of cluster-by-item interaction variance across our empirical datasets leads us to conjecture that similar results would obtain in other contexts, 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. We did not emphasize another important issue with VAMs here, namely, that the pretest scores used as covariates themselves contain 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. These issues have further implications for the use of cluster-mean pretest scores in the Mundlak model (Asparouhov & Muthén, 2019; Hamaker & Muthén, 2020). 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 using `brms` may provide an alternative (Bürkner, 2017). However, replication of our simulation study in Appendix C with varying degrees of pretest measurement error shows that the pattern of results is unchanged.

3. Our reliability estimates treat the variance components as known, when they are estimated with uncertainty. Thus, when the number of clusters and/or items is low, estimates of σ^2_ν 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 Generalizability Theory 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; 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^2_ν 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., 2024, 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, 2024b; Gilbert, Zhang, et al., 2025).

5. Our models assume a unidimensional construct, which is one reason we estimate separate models for the different outcomes from the subset of studies that contributed 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). As we show in Appendix D, using our proposed item interaction model can capture such variation even when data from only a single time point is available. 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/https://doi.org/10.1080/19345747.2024.2361337

Amrein-Beardsley, A. (2014, April). *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

Andrabi, T., Das, J., Ijaz Khwaja, A., & Zajonc, T. (2011). Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics*, *3*(3), 29–54. https://doi.org/10.1257/app.3.3.29

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.

Aslantas, I. (2020). Impact of Contextual Predictors on Value-Added Teacher Effectiveness Estimates. *Education Sciences*, *10*(12), 390. https://doi.org/10.3390/educsci10120390

Asparouhov, T., & Muthén, B. (2019). Latent Variable Centering of Predictors and Mediators in Multilevel and Time-Series Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Association, A. E. R. (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

Atteberry, A., & Mangan, D. (2020). The Sensitivity of Teacher Value-Added Scores to the Use of Fall or Spring Test Scores. *Educational Researcher*, *49*(5), 335–349. https://doi.org/10.3102/0013189X20922993

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

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

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). https://doi.org/10.18637/jss.v067.i01

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

Berry, J., Karlan, D., & Pradhan, M. (2018). The impact of financial education for youth in Ghana. *World Development*, *102*, 71–89.

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

Bollen, K. A. (1989). *Structural Equations with Latent Variables* (1st ed.). Wiley. https://doi.org/10.1002/9781118619179

Brandt, K. (2023). When Private Beats Public: A Flexible Value-Added Model with Tanzanian School Switchers. *Economic Development and Cultural Change*, *72*(1), 159–206. https://doi.org/10.1086/718893

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. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

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

Bruhn, M., de Souza Leão, L., Legovini, A., Marchetti, R., & Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*, *8*(4), 256–295.

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). 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., & Ho, A. D. (2013). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 190–215. https://doi.org/10.3102/1076998611435413

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

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*, 533–559.

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

de Barros, A., Fajardo-Gonzalez, J., Glewwe, P., & Sankar, A. (2024). The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling. *The Economic Journal*, *134*(659), 959–984.

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

Dimitrov, D. M. (2003). Marginal True-Score Measures and Reliability for Binary Items as a Function of Their IRT Parameters. *Applied Psychological Measurement*, *27*(6), 440–458. https://doi.org/10.1177/0146621603258786

Domingue, B., 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

Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental Evidence on Four Policies to Increase Learning at Scale. *The Economic Journal*, ueae003.

Duflo, E., Berry, J., Mukerji, S., & Shotland, M. (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. *3ie Impact Evaluation Report*, *22*.

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

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri. *Statistics and Public Policy*, *1*(1), 19–27. https://doi.org/10.1080/2330443X.2013.856152

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. (2024a). How measurement affects causal inference: Attenuation bias is (usually) more important than scoring weights. *Ed Working Papers*. https://edworkingpapers.com/ai23-766

Gilbert, J. B. (2024b). 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.

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2024). Estimating Heterogeneous Treatment Effects with Item-Level Outcome Data: Insights from Item Response Theory. https://doi.org/10.48550/ARXIV.2405.00161

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.

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2024). Leveraging item parameter drift to assess transfer effects in vocabulary learning. *Applied Measurement in Education*, *37*(3), 240–257. https://doi.org/https://doi.org/10.1080/08957347.2024.2386934

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/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., 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

Glatz, T., Tops, W., Borleffs, E., Richardson, U., Maurits, N., Desoete, A., & Maassen, B. (2023). Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers: A cluster randomised controlled trial. *PeerJ*, *11*, e15499.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*(4), 395–418.

Goldhaber, D., & Hansen, M. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, *80*(319), 589–612. https://doi.org/10.1111/ecca.12002

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

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

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. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

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. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

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

Jackson, C. D. (2013). *Modeling teacher effectiveness as a function of student ability* [Master's thesis, University of Texas]. https://repositories.lib.utexas.edu/items/04ed0c01-8343-49f1-b04a-9417ae9313d8/full

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., & 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

Kersting, N. B., Chen, M.-K., & Stigler, J. W. (2013). Value-added Teacher Estimates as Part of Teacher Evaluations: Exploring the Effects of Data and Model Specifications on the Stability of Teacher Value-added Scores. *Education Policy Analysis Archives*, *21*, 7. https://doi.org/10.14507/epaa.v21n7.2013

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.

Kinsler, J. (2012). Beyond Levels and Growth: Estimating Teacher Value-Added and its Persistence. *Journal of Human Resources*, *47*(3), 722–753. https://doi.org/10.3368/jhr.47.3.722

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. 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

Konstantopoulos, S., & Chung, V. (2011). The Persistence of Teacher Effects in Elementary Grades. *American Educational Research Journal*, *48*(2), 361–386. https://doi.org/10.3102/0002831210382888

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

Llauradó, E., Tarro, L., Moriña, D., Queral, R., Giralt, M., & Solà, R. (2014). EdAl-2 (Educacio en Alimentacio) programme: Reproducibility of a cluster randomised, interventional, primary-school-based study to induce healthier lifestyle activities in children. *BMJ Open*, *4*(11), e005496.

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. (2009). Exploring Student-Teacher Interactions in Longitudinal Achievement Data. *Education Finance and Policy*, *4*(4), 439–467. https://doi.org/10.1162/edfp.2009.4.4.439

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. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics

Achievement Measures. *Journal of Educational Measurement*, *44*(1), 47–67. https://doi.org/10.1111/j.1745-3984.2007.00026.x

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.

Loeb, S., Soland, J., & Fox, L. (2014). Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers With Their English Learners and Non-English Learners. *Educational Evaluation and Policy Analysis*, *36*(4), 457–475. https://doi.org/10.3102/0162373714527788

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

Maruyama, T. (2022). Strengthening Support of Teachers for Students to Improve Learning Outcomes in Mathematics: Empirical Evidence on a Structured Pedagogy Program in El Salvador. *International Journal of Educational Research*, *115*, 101977.

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.

Mohohlwane, N., Taylor, S., Cilliers, J., & Fleisch, B. (2023). Reading Skills Transfer Best from Home Language to a Second Language: Policy Lessons from Two Field Experiments in South Africa. *Journal of Research on Educational Effectiveness*, 1–24. https://doi.org/https://doi.org/10.1080/19345747.2023.2279123

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

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.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Education Policy Analysis Archives*, *18*, 23. https://doi.org/10.14507/epaa.v18n23.2010

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. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163–193. https://doi.org/10.3102/0002831210362589

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.

Persson, M., Andersson, K., Zetterberg, P., Ekman, J., & Lundin, S. (2020). Does deliberative education increase civic competence? Results from a field experiment. *Journal of Experimental Political Science*, *7*(3), 199–208.

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.

Pollack, J. M., Rock, D. A., Weiss, M. J., & Atkins-Burnett, S. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (tech. rep. No. (NCES 2005–062). National Center for Education Statistics. Washington, DC. https://nces.ed.gov/pubs2005/2005062.pdf

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. 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

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

Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review*, *110*(2), 364–400.

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 (C. Sutherland, Ed.). *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

Schreinemachers, P., Baliki, G., Shrestha, R. M., Bhattarai, D. R., Gautam, I. P., Ghimire, P. L., Subedi, B. P., & Brück, T. (2020). Nudging children toward healthier food choices: An experiment combining school and home gardens. *Global Food Security*, *26*, 100454.

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

Sebele, M., Jeffery, M., Mwanza, M., Mwai, L., & Rudasingwa, M. (2023). Improving reading proficiency in early childhood education classrooms: Evidence from Liberia. https://doi.org/https://riseprogramme.org/sites/default/files/inline-files/Rudasingwa_Improving_Reading_Proficiency_ECE_Liberia.pdf

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

Shores, K. A., & Student, S. R. (2024). Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory. https://doi.org/10.26300/48D6-MX29

Sørensen, Ø. (2024). Multilevel Semiparametric Latent Variable Modeling in R with "galamm". *Multivariate Behavioral Research*. https://doi.org/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

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36. https://doi.org/10.3102/10769986029001011

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

Van De Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement*, *20*(2), 269–285. https://doi.org/10.1080/09243450902883946

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

Woods-Townsend, K., Hardy-Johnson, P., Bagust, L., Barker, M., Davey, H., Griffiths, J., Grace, M., Lawrence, W., Lovelock, D., Hanson, M., et al. (2021). A cluster-randomised controlled trial of the LifeLab education intervention to improve health literacy in adolescents. *PLoS One*, *16*(5), e0250545.

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

Yeh, S. S. (2012). The Reliability, Impact, and Cost-Effectiveness of Value-Added Teacher Assessment Methods. *Journal of Education Finance*, *37*(4), 374–399. https://doi.org/10.1353/jef.2012.a475491

Zhao, V. Y., Hilgendorf, D., Yoshikawa, H., & Michael, D. (2023). Impacts of Ahlan Simsim TV Program in Pre-Primary Classrooms in Jordan on Children's Emotional Development: A Randomized Controlled Trial.

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}:

$$y_{ijk} = u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk} \tag{22}$$

$$u_k \sim N(0, \sigma_u^2) \tag{23}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{24}$$

$$b_i \sim N(0, \sigma_b^2) \tag{25}$$

$$\nu_{ik} \sim N(0, \sigma_\nu^2) \tag{26}$$

$$e_{ijk} \sim N(0, \sigma_e^2), \tag{27}$$

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 23, we fit the following misspecified model that omits the cluster-by-item interactions ν_{ik}:

$$y_{ijk} = u_k + \theta_{jk} + b_i + e_{ijk} \tag{28}$$

$$u_k \sim N(0, \sigma_u^2) \tag{29}$$

$$\theta_{jk} \sim N(0, \sigma_\theta^2) \tag{30}$$

$$b_i \sim N(0, \sigma_b^2) \tag{31}$$

$$e_{ijk} \sim N(0, \sigma_e^2). \tag{32}$$

We first establish ANOVA estimators for the variance components in Equation 28. 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:

$$SS_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{.jk} - \bar{y}_{i..} + \bar{y}_{...})^2 \tag{33}$$

$$SS_i = JK\sum_{i=1}^{I}(\bar{y}_{i..} - \bar{y}_{...})^2 \tag{34}$$

$$SS_{j|k} = I\sum_{k=1}^{K}\sum_{j=1}^{J}(\bar{y}_{.jk} - \bar{y}_{..k})^2 \tag{35}$$

$$SS_k = IJ\sum_{k=1}^{K}(\bar{y}_{..k} - \bar{y}_{...})^2. \tag{36}$$

Using these, we define the mean squares:

$$MS_e = \frac{SS_e}{IJK} \tag{37}$$

$$MS_i = \frac{SS_i}{I-1} \tag{38}$$

$$MS_{j|k} = \frac{SS_{j|k}}{K(J-1)} \tag{39}$$

$$MS_k = \frac{SS_k}{K-1}. \tag{40}$$

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

$$\widehat{\sigma}_e^2 = MS_e \tag{41}$$

$$\widehat{\sigma}_\theta^2 = \frac{MS_{j|k} - MS_e}{I} \tag{42}$$

$$\widehat{\sigma}_u^2 = \frac{MS_k - MS_{j|k}}{IJ}. \tag{43}$$

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 28 when Equation 23 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 $y_{...}$.

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

Then we regroup this expression as sums of each parameter.

$$\overline{y}_{..k} = \big(\tfrac{1}{I}\sum_{i=1}^{I} b_i\big) + \big(\tfrac{1}{J}\sum_{j=1}^{J}\theta_{jk}\big) + u_k + \big(\tfrac{1}{I}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ijk}\big). \tag{45}$$

Similarly for the overall (grand) sample mean:

$$\begin{aligned}
\overline{y}_{...} &= \frac{1}{IJK}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijk} && (46)\\
&= \frac{1}{IJK}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}\big(b_i + \theta_{jk} + u_k + \nu_{ik} + e_{ijk}\big), && (47)\\
&= \big(\tfrac{1}{I}\sum_{i=1}^{I} b_i\big) + \big(\tfrac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J}\theta_{jk}\big) + && (48)\\
&\big(\tfrac{1}{K}\sum_{k=1}^{K} u_k\big) + \big(\tfrac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJK}\sum_{k,i,j} e_{ijk}\big). && (49)
\end{aligned}$$

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

$$\begin{aligned}
\big(\overline{y}_{..k} - \overline{y}_{...}\big) &= \big(\tfrac{1}{J}\sum_{j=1}^{J}\theta_{jk} - \tfrac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J}\theta_{jk}\big) + \big(u_k - \tfrac{1}{K}\sum_{k=1}^{K} u_k\big) + && (50)\\
&\big(\tfrac{1}{I}\sum_{i=1}^{I}\nu_{ik} - \tfrac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\nu_{ik}\big) + \big(\tfrac{1}{IJ}\sum_{i,j} e_{ijk} - \tfrac{1}{IJK}\sum_{k,i,j} e_{ijk}\big). && (51)
\end{aligned}$$

This is equivalent to

$$\overline{y}_{..k} - \overline{y}_{...} = \left(\overline{\theta}_{.k} - \overline{\theta}_{..}\right) + \left(u_k - \overline{u}\right) + \left(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\right) + \left(\overline{e}_{..k} - \overline{e}_{...}\right). \tag{52}$$

Now recall that

$$SS_k = IJ\sum_{k=1}^{K}\Big[\left(\overline{y}_{..k} - \overline{y}_{...}\right)\Big]^2, \quad MS_k = \frac{SS_k}{K-1}, \tag{53}$$

and notice that

$$(\overline{y}_{..k} - \overline{y}_{...})^2 = (\overline{\theta}_{.k} - \overline{\theta}_{..})^2 + (u_k - \overline{u})^2 + \left(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\right)^2 + \tag{54}$$

$$\left(\overline{e}_{..k} - \overline{e}_{...}\right)^2 + \text{(cross terms)}, \tag{55}$$

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

$$\mathbb{E}[(\overline{\theta}_{.k} - \overline{\theta}_{..})^2] = \frac{(K-1)\sigma_\theta^2}{J} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}[(\overline{\theta}_{.k} - \overline{\theta}_{..})^2] = I\sigma_\theta^2 \tag{56}$$

$$\mathbb{E}[(u_k - \overline{u})^2] = (K-1)\sigma_u^2 \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}[(u_k - \overline{u})^2] = IJ\sigma_u^2 \tag{57}$$

$$\mathbb{E}\left[\left(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\right)^2\right] = \frac{(K-1)\sigma_\nu^2}{I} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}\left[\left(\tfrac{1}{I}\sum_i \nu_{ik} - \overline{\nu}\right)^2\right] = J\sigma_\nu^2 \tag{58}$$

$$\mathbb{E}\Big[(\overline{e}_{..k} - \overline{e}_{...})^2\Big] = \frac{(K-1)\sigma_e^2}{IJ} \Rightarrow IJ\frac{1}{K-1}\sum_{k=1}^{K}\mathbb{E}\Big[(\overline{e}_{..k} - \overline{e}_{...})^2\Big] = \sigma_e^2, \tag{59}$$

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

$$\mathbb{E}[MS_k] \;=\; IJ\sigma_u^2 \;+\; I\sigma_\theta^2 \;+\; J\sigma_\nu^2 \;+\; \sigma_e^2. \tag{60}$$

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

$$\mathbb{E}[MS_{j|k}] = I\sigma_\theta^2 + \sigma_e^2 \tag{61}$$

$$\mathbb{E}[MS_k] = IJ\sigma_u^2 \;+\; I\sigma_\theta^2 \;+\; J\sigma_\nu^2 + \sigma_e^2 \tag{62}$$

$$\mathbb{E}[MS_i] = JK\sigma_b^2 \;+\; J\sigma_\nu^2 \;+\; \sigma_e^2 \tag{63}$$

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

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

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

$$\mathbb{E}[\widehat{\sigma}_e^2] = \mathbb{E}[MS_e] = \sigma_e^2 + \alpha\sigma_\nu^2 \tag{65}$$

$$\mathbb{E}[\widehat{\sigma}_\theta^2] = \mathbb{E}\left[\frac{MS_{j|k} - MS_e}{I}\right] = \sigma_\theta^2 - \frac{\alpha\sigma_\nu^2}{I} \tag{66}$$

$$\mathbb{E}[\widehat{\sigma}_u^2] = \mathbb{E}[\frac{MS_k - MS_{j|k}}{IJ}] = \sigma_u^2 + \frac{\sigma_\nu^2}{I}. \tag{67}$$

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

$$\widehat{\rho} = \frac{\widehat{\sigma_u^2}}{\widehat{\sigma_u^2} + \frac{\widehat{\sigma_\theta^2}}{J} + \frac{\widehat{\sigma_e^2}}{IJ}} \tag{68}$$

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

$$\mathbb{E}[\widehat{\rho}] \approx \frac{\mathbb{E}[\widehat{\sigma_u^2}]}{\mathbb{E}[\widehat{\sigma_u^2}] + \mathbb{E}[\frac{\widehat{\sigma_\theta^2}}{J}] + \mathbb{E}[\frac{\widehat{\sigma_e^2}}{IJ}]} \tag{69}$$

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

$$= \frac{\sigma_u^2 + \frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{71}$$

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:

$$\mathbb{E}[\widehat{\rho}] - \rho \approx \frac{\frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}} \tag{72}$$

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:

$$y_{ijk} = u_k + \theta_{jk} + b_i + \nu_{ik} + e_{ijk} \tag{73}$$

$$V(y_{ijk}) = \sigma_u^2 + \sigma_\theta^2 + \sigma_b^2 + \sigma_\nu^2 + \sigma_e^2. \tag{74}$$

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:

$$V(\text{post}_{jk}) = \sigma_u^2 + \sigma_\theta^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_e^2}{I}. \tag{75}$$

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:

$$\mathbb{E}[\widehat{\sigma_u^2}] = \sigma_u^2 + \frac{\sigma_\nu^2}{I} \tag{76}$$

$$\mathbb{E}[\widehat{\sigma_\theta^2}] = \sigma_\theta^2 + \frac{\sigma_e^2}{I}. \tag{77}$$

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

$$\widehat{\rho}_{\text{mean}} = \frac{\widehat{\sigma_u^2}}{\widehat{\sigma_u^2} + \frac{\widehat{\sigma_\theta^2}}{J}} \tag{78}$$

$$= \frac{\sigma_u^2 + \frac{\sigma_\nu^2}{I}}{\sigma_u^2 + \frac{\sigma_\nu^2}{I} + \frac{\sigma_\theta^2}{J} + \frac{\sigma_e^2}{IJ}}. \tag{79}$$

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 Simulation Design and Results

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. 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 C.1: Fixed and Varying Simulation Design Factors

Simulation Factor	Notation	Values
Number of Subjects	JK	1,000
Number of Clusters	K	50
Number of Items	I	5, 20, 50
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 to 1 in increments of .05
Student Residual Variance	σ_θ^2	1
Pretest Coefficient	β_1	1
Cluster Covariate Coefficient	γ_1	.5
Intraclass Correlation on the Pretest	ICC_{pre}	0, .25, .5
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 $N \sim (0, 1)$.

Table C.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 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 .05. The other varying simulation factors include the number of items and the intraclass correlation (ICC) on the pretest variable. We vary the number of items at 5, 20, and 50 to represent short, moderate, and long assessments, and we vary the pretest ICC at 0, .25, and .5 to represent no sorting, moderate sorting, and high sorting of students within clusters (Hedges & Hedberg, 2007). We fix the remaining values to those specified in Table C.1 and perform 60 replicates of each condition.

We generate the data and fit two models, one assuming constant VA effects across items, the other allowing for cluster-by-item interactions. Our primary goal is to examine estimates of the reliability of u_k derived from each model and to determine how these estimates compare to the analytic solution described earlier. We expect that the reliability results from the simulation will 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).

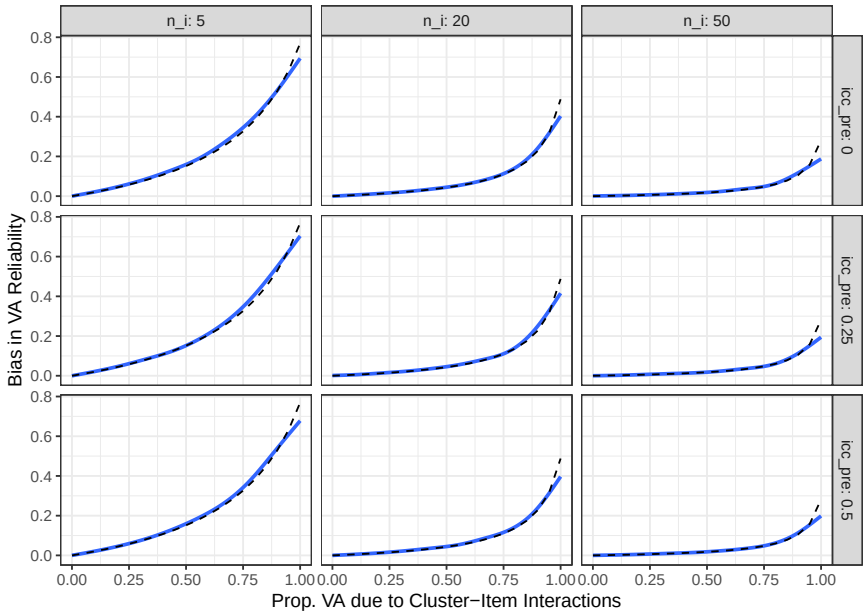
Figure C.1 shows the simulation results. The x-axis shows the proportion of VA variance attributable to cluster-by-item interactions ($\frac{\sigma_\nu^2}{\sigma_u^2+\sigma_\nu^2}$) and the y-axis shows the estimated bias in VA reliability (the misspecified model minus the correct model). The blue lines show LOESS curves fit to the simulation results 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 near-perfect 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 ICC for the pretest scores does not affect the results, suggesting that this pattern of results is robust to sorting among clusters fully captured by the covariates.

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 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 replicate our results using dichotomous items and a pretest with varying degrees of measurement error and find that the pattern of results is unchanged.

As an additional sensitivity check, we 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.

Figure C.1: VAMs Assuming Constant Effects Across Items Overestimate Reliability



The y-axis shows the estimated bias 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 ($\frac{\sigma^2_\nu}{\sigma^2_u+\sigma^2_\nu}$). The black dashed lines represent the theoretical prediction of the reliability bias derived from the formula in Appendix A. The blue lines represent LOESS curves fit to the simulation results. n_i = number of items; icc_pre = intraclass correlation on the pretest variable.

D Equivalence of G-Theory and Correlation Estimates of Reliability in the Tanzania Data

To further illustrate the interpretation and affordances of the item-level VAMs, we use the large nature of the case study dataset to demonstrate the equivalence between the reliabilities derived from the multilevel model with the corresponding correlations between replications. We approach this demonstration in a bootstrap resampling framework using the following steps:

1. Create a perfectly balanced subsample by limiting the data to exactly 50 randomly sampled students per school (excluding schools with fewer than 50 students) and randomly sample 100 schools. Remove a single item so that we have an even number of items (8 items total), allowing for an equal split between samples. (We remove Swahili because it has the lowest item discrimination of the 9 items in the full sample.)

2. Split the data into a training set and two test sets with 25 students and 4 items each. The training set and the test sets contain different samples of students. The first test set contains the same 4 items as the training set and the second test set contains the other 4 items (the initial 4 items of the training set are discarded).

3. Fit both Equations 20 and 21 to the training set, extract the VA estimates for each school, and calculate the model-implied VA reliability.

4. Fit Equation 20 to both test sets and extract the VA estimates.

5. Correlate the VA estimates from the training set to those of the test sets.

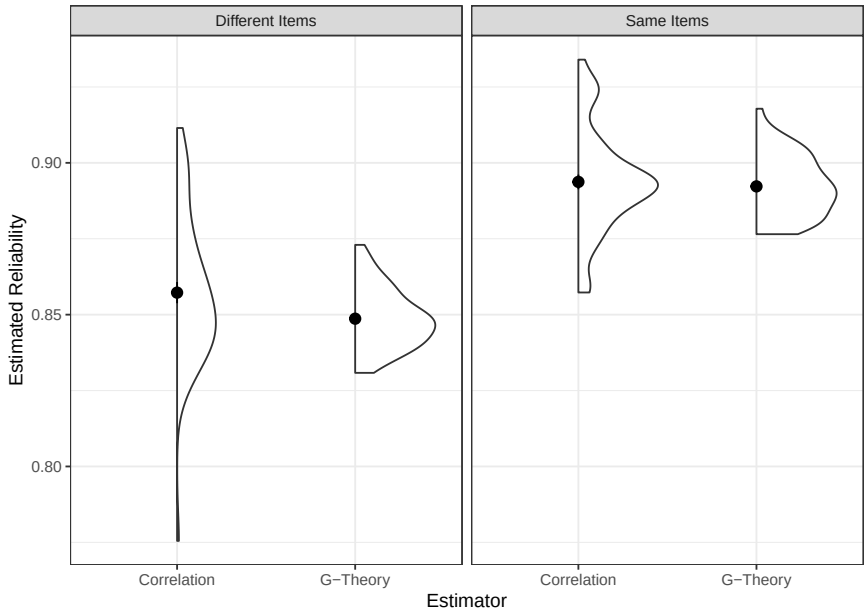
6. Repeat steps 2-5 1,000 times and collect the results for analysis.

This bootstrap resampling procedure allows us to determine the extent to which the model-based estimated reliabilities accurately match the implied correlations between replications when only students vary *or* both students and items vary between replications. Theoretically, these two estimators of reliability should coincide in expectation.

Figure D.1 shows our bootstrap resampling approach comparing the model-based reliabilities to correlation-based reliabilities. In line with our derivations, the estimates from both methods are similar. That is, when students differ but items are held constant, the reliability estimates from Model 2 match the correlation between replications, on average. When both students and items differ, the estimates of Model 3 match the correlation between replications, on average. Furthermore, the reliability estimates from the Generalizability Theory formulas are more precisely estimated (SDs $\approx .01$) than those derived from direct correlations between replications (SDs $\approx .02$),

suggesting that the G-Theory estimator is more efficient, at least in these data. Thus, even if only one time point is available, researchers can use our approach to precisely estimate the reliability of a VA estimate *had different items been administered*, even when the number of items is small.

Figure D.1: Comparison of Reliability Estimators in Brandt (2023) Data



The y-axis shows the estimated reliability of a school VA estimate and the x-axis shows the reliability estimator, either a correlation between replications or the Generalizability Theory/multilevel model formula. The panels show whether the same items or different items are selected. The analyses derive from a balanced subsample of the full data from Brandt (2023). Each resampled dataset contains 100 schools, 25 students, and 4 items. The distributions show the results of 1,000 replications.