Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1226

The Sensitivity of Value-Added Estimates to Test Scoring Decisions

Joshua B. Gilbert
Harvard University

James G. Soland
University of Virginia

Benjamin W. Domingue
Stanford University

Value-Added Models (VAMs) are both common and controversial in education policy and accountability research. While the sensitivity of VAMs to model specification and covariate selection is well documented, the extent to which test scoring methods (e.g., mean scores vs. IRT-based scores) may affect VA estimates is less studied. We examine the sensitivity of VA estimates to scoring method using empirical item response data from 18 education datasets. We show that VA estimates are frequently highly sensitive to scoring method, holding constant students and items. While the various test scores are highly correlated, on average, different scoring approaches result in VA percentile ranks that vary by over 20 points, and over 50% of teachers or schools are ranked in more than one quartile of the VA distribution. Dispersion in VA ranks is reduced with more complete item response data. We conclude that consideration of both measurement error and model uncertainty is necessary for appropriate interpretation of VAMs.

VERSION: September 2025

Suggested citation: Gilbert, Joshua B., James G. Soland, and Benjamin W. Domingue. (2025). The Sensitivity of Value-Added Estimates to Test Scoring Decisions. (EdWorkingPaper: 25-1226). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/g4gn-s810

The Sensitivity of Value-Added Estimates to Test Scoring Decisions

Joshua B. Gilbert [1], James G. Soland [2], and Benjamin W. Domingue [3]

[1]Harvard Graduate School of Education
[2]University of Virginia School of Education and Human Development
[3]Stanford University Graduate School of Education

September 1, 2025

Abstract

Value-Added Models (VAMs) are both common and controversial in education policy and accountability research. While the sensitivity of VAMs to model specification and covariate selection is well documented, the extent to which test scoring methods (e.g., mean scores vs. IRT-based scores) may affect VA estimates is less studied. We examine the sensitivity of VA estimates to scoring method using empirical item response data from 18 education datasets. We show that VA estimates are frequently highly sensitive to scoring method, holding constant students and items. While the various test scores are highly correlated, on average, different scoring approaches result in VA percentile ranks that vary by over 20 points, and over 50% of teachers or schools are ranked in more than one quartile of the VA distribution. Dispersion in VA ranks is reduced with more complete item response data. We conclude that consideration of both measurement error and model uncertainty is necessary for appropriate interpretation of VAMs.

Keywords: value-added model, item response theory, test scoring, reliability, education policy

Corresponding author: joshua_gilbert@g.harvard.edu

The authors report no conflicts of interest.

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

Data and Code Availability: All datasets are publicly available in the Item Response Warehouse (IRW) (Domingue, Braginsky, et al., 2025). See Gilbert, Himmelsbach, Soland, et al. (2025) and Gilbert, Himmelsbach, Miratrix, et al. (2025) for more detail on these datasets. Our replication materials will be available at the following URL upon publication: https://doi.org/10.7910/DVN/AJDUN2.

1 Introduction

Value-added models (VAMs) attempt to estimate the causal effects of teachers and schools on student learning, most commonly by adjusting year end test scores for prior year test scores in regression models. The use of VAMs has inspired considerable debate among education researchers, particularly in the context of high-stakes decisions based on VA estimates (American Educational Research Association, 2015; Amrein-Beardsley, 2014; Amrein-Beardsley et al., 2016; Harris, 2009). The methodological approaches to estimating VAMs have thus received extensive commentary in areas such as model specification (e.g., covariate selection, functional form, covariate measurement error, the use of multiple pretests or non-test score covariates), psychometric considerations (e.g., the effects of rapid guessing, and heterogeneous effects across outcome items), and idiosyncratic issues such as the timing of a test within a school year (Atteberry & Mangan, 2020; Bitler et al., 2021; Chetty, Friedman, & Rockoff, 2014; Gilbert, Himmelsbach, Miratrix, et al., 2025; Jensen et al., 2018; Levy et al., 2019, 2023; Papay, 2011).

One methodological consideration for VAMs that has remained relatively unexplored has been the method in which the test is scored (Briggs & Weeks, 2009; Ng & Koretz, 2015). That is, in cases in which the outcomes in VAMs are test scores, such scores are constructed from student responses to individual test items. Many options exist for constructing an overall score from these responses; options include the familiar sum or mean scores from Classical Test Theory (CTT) as well as more complex approaches from Item Response Theory (IRT). Scoring decisions include whether to use an explicit measurement model, which measurement model to use, how to estimate item parameters, and how to produce scores based on that model (Soland et al., 2024). While test scores derived from different scoring models are generally highly correlated, some evidence suggests that substantive inferences can be highly sensitive to the choice of scoring system (McNeish, 2022; Soland, 2022; Soland et al., 2024).

In this study, we examine the impact of the scoring model on the properties of VAMs, principally, the consistency of individual school or teacher VA estimates. Using 18 empirical datasets in education with item-level baseline and endline data in which students are clustered within classrooms or schools, we show that VA estimates are often quite sensitive to test scoring decisions. Across the empirical datasets, we find that changes only to the scoring method can shift VA percentile ranks by an average of 20 points, 50% of units score across multiple quartiles, and that accounting for model uncertainty inflates VA standard errors by about 10%, holding students and items constant. The dispersion of VA estimates is reduced when all respondents answer all items (as opposed to designs where each respondent answers a subset of items). Our results highlight the importance of measurement considerations in educational policy and accountability.

2 Background

2.1 Decisions Involved in Scoring Item Responses

At first glance, the primary decision in "scoring" is whether to use a sum score versus an IRT model. However, we argue that there are several steps involved in producing scores; we colloquially refer to this entire process as "scoring." Following the taxonomy of Soland et al. (2024), these decision steps include (1) whether to use an explicit measurement model, (2) which measurement model to use, (3) how to estimate item parameters, and (4) how to produce scores based on that model. We briefly relate each of those decisions back to uses of VAM, including how these decisions could relate to VAM sensitivity to specification.

2.1.1 Step 1. Whether to Use an Explicit Measurement Model

The first decision involves whether to use a measurement model versus a sum score. Researchers often like to use sum scores (variously referred to as total scores or mean scores) because they are intuitive, easy to explain and, seemingly, do not include many assumptions (Sijtsma et al., 2024a). Yet, as McNeish and Wolf (2020) (among others) point out, one way to view sum scores is that they involve an implicit—and highly constrained—measurement model. In fact, by imposing constraints on a measurement model (e.g., equal item weights/factor loadings), one can produce factor scores that are perfectly correlated with sum scores. While sum scores can seem like they are relatively assumption free, they may rely on stringent assumptions that are frequently not met in practice under a latent variable view of measurement (McNeish, 2022; McNeish & Wolf, 2020).[1]

Further, violations of the assumptions undergirding the sum score approach can lead to major problems. For example, true rank orderings of individuals can be extremely distorted (McNeish & Wolf, 2020). In the context of gains or growth, estimates of change over time and variability in that change can be severely understated (Bauer & Curran, 2016; Gorter et al., 2020; Kuhfeld & Soland, 2022). When estimating group differences in change, such as in the case of randomized control trials and quasi-experimental designs, results can also be distorted (Gorter et al., 2020; Soland, 2022, 2024; Soland, Edwards, & Talbert, 2025; Soland et al., 2024). Using sum scores can also create problems in clustered, longitudinal designs (Kuhfeld & Soland, 2022)—that is, in a scenario that directly parallels the VAM specification described in Section 2.2. Despite these issues, sum scores remain popular in education and psychology (McNeish & Wolf, 2020), though to our knowledge are rarely applied in VAM contexts such as US state accountability systems where IRT-based scores are commonplace.

2.1.2 Step 2. Which Measurement Model to Use

Once a researcher decides to use an explicit measurement model, there are many possible measurement models (e.g., IRT models) that can be fit to the item responses. One choice related to measurement model specification is, regardless of the number of latent variables, how to parameterize the item parameters. Consider a test comprised of I dichotomously scored items answered by student j in classroom or school k, with item responses denoted y_{ijk}. Under CTT, the sum (or mean) of the item responses forms the total test score: $\text{post}_{jk} = \Sigma_i y_{ijk}$. Alternatively, IRT approaches explicitly model the probability of a correct response as a function of person ability θ_j and item

[1]Given our emphasis on IRT models, we take a latent variable perspective on measurement of student achievement for the purposes of this study (Borsboom, 2005). When a reflective latent variable model is not assumed, sum scores can still be considered meaningful summaries of performance under less restrictive assumptions, as in Generalizability Theory or network psychometrics frameworks. See Markus and Borsboom (2024) for more discussion of these issues.

parameters a_i (discrimination), b_i (difficulty), and c_i (pseudo-guessing):

$$\Pr(y_{ijk} = 1) = c_i + (1 - c_i)\text{logit}^{-1}[a_i(\theta_j - b_i)]. \tag{1}$$

IRT models such as the three-parameter logistic (3PL) model above use a measurement model with these parameters to generate latent trait estimates, denoted $\widehat{\theta}_{jk}$, by finding the score that maximizes the probability of a student's observed item responses, given the item parameters, thus maximizing the information contained in the score (Embretson & Reise, 2000; Jessen et al., 2018; Rhemtulla & Savalei, 2025). Alternative IRT models for dichotomous responses include the 2PL, in which all $c_i = 0$, and the 1PL or Rasch model, in which all $c_i = 0$ and $a_i = a$.

IRT-based scores are generally argued to have superior properties to CTT-based scores in single timepoint and longitudinal contexts. Specific to VAMs, another reason is that, when using a 1PL or Rasch model, IRT scores possess properties that are necessary (but not sufficient) to place scores on an interval rather than ordinal scale when model assumptions are met (Briggs & Weeks, 2009; Camilli, 2018; Gilbert, 2025b) (though some have explored ordinal models for VAMs to relax this assumption; see Ballou, 2009). In the VAM context, Soland (2017) shows that even mild changes to the scoring scale (and, specifically, departures from the equal interval assumption) can substantially shift the odds of a teacher being designated as high or low performing. Related work shows that VAMs can also be sensitive to vertical scaling decisions involved in measuring learning across broad domains and multiple years of development (Martineau, 2006).

Another choice related to measurement model specification is what the factor structure should be, especially in the longitudinal context, and whether to build multiple group membership into the model. Both have been shown to have large impacts on estimates from randomized controlled trials (RCTs), latent growth curve models, and growth mixture models (Edwards & Soland, 2024; Kuhfeld & Soland, 2022; McNeish, 2022; McNeish & Wolf, 2020; Soland, 2021; Soland, Cole, et al., 2025; Soland et al., 2024). For example, consider the case of a pre/post RCT. One could fit a unidimensional, single group IRT model. Such a model implicitly assumes that all groups have the same mean and variance at all timepoints, an assumption that lacks face validity. By contrast, a multigroup and multi-timepoint (that is, multidimensional with one factor per timepoint) model allows each group and timepoint to have its own mean and variance in the model, matching the study design. In a VAM design, there are by definition multiple timepoints and, depending on how one looks at it, multiple groups.

A final, arguably more minor decision in this step that relates to our investigation is what item response model to use once the item parameters and factor structure have been solidified. For instance, for polytomous items, options include the Partial Credit Model (PCM), Generalized PCM (GPCM), and Graded Response Model (GRM), whose properties and conceptions of order differ in subtle but important ways (Domingue, Kanopka, et al., 2025; Nalbandyan et al., 2024). While such decisions tend to be less consequential relative to questions of factor structure and whether the items should have different weights, they can nonetheless impact inferences based on resulting scores.

2.1.3 Step 3. How to Estimate Item Parameters

Another step in the process relates to how to estimate parameters. Generally, this step is not a major part of our investigation, therefore we discuss it only briefly. Essentially, one needs to decide on an approach to estimate the item parameters (e.g., full information maximum likelihood [FIML] versus Markov Chain Monte Carlo [MCMC]). Once the model parameters are estimated, they are considered "calibrated" and used to score the measure (Step 4). As part of this step, one needs to decide whether the sample size is sufficient to accurately estimate those parameters. For example, complex IRT models such as the 3PL would likely not be appropriate with fewer than several hundred examinees (Cuhadar, 2022; Domingue et al., 2024).

2.1.4 Step 4. How to Produce Scores Based on the Measurement Model

The final step is determining how to score the item responses using the calibrated parameters. One option is maximum likelihood estimation (MLE), which has many desirable statistical properties (Baker, 2001). However, MLE scores are undefined when a person's item responses all fall in the top or bottom category. While this is not generally a problem on achievement tests (except ones with floor and ceiling effects where some respondents get all of the items wrong or right), surveys are often a different matter, with individuals frequently using only top or bottom response categories (Edwards & Soland, 2024). A workaround is to use empirical Bayes approaches like Expected a Posteriori (EAP) scoring. In this case, use of a prior helps avoid undefined scores via the posterior. A key decision with EAP scoring is which means and variances (estimated during calibration then treated as known during scoring) to use as priors. Scores that are noisy are shrunk to those priors. If a unidimensional model is used, then only one mean and variance provide the basis for shrinkage. If a multidimensional or multigroup model is used, then scores are shrunk towards group- and time-specific means (Briggs, 2008).[2]

The implications here become clearer in a concrete example like an RCT. In an RCT, using a unidimensional IRT model with EAP scoring results in scores being shrunk towards a single mean that smooths over control/treatment and pre/post differences. By contrast, using a multigroup multi-timepoint model means scores are shrunk towards pre/post and control/treatment specific means. In simpler terms, one approach pulls control and treatment groups together, the other pulls them apart. Although the risk-averse researcher might opt to pull those scores together (unidimensional), doing so often leads to downwardly biased treatment effect estimates and high Type II error rates (Soland et al., 2024). In contrast, using the multigroup multi-timepoint model addresses these issues, usually without an increase in Type I errors. Type I errors are not impacted (except with extremely short sample sizes or measures with fewer than five items) because allowing different groups (here control and treatment) to have separate means in the model does not necessarily result in the groups having different means. That is, both will have different means in the model, but those means could both be zero—they would just be estimated separately.

The extent to which scoring decisions—both the last scoring step and the prior three steps—may affect VAMs is less studied, with a few exceptions. For example, Briggs and Weeks (2009) examine

[2]Alternative scoring approaches such as weighted likelihood estimation (WLE) are available, but less commonly used (Embretson & Reise, 2000; Warm, 1989). As such, we do not explore additional scoring systems for the purposes of this study.

different approaches to vertical scaling in Colorado schools, and find that the ordering of VA estimates is relatively insensitive to the scaling choice, though precision is affected. Most similar to the present study, Ng and Koretz (2015) examine two scaling approaches in New York middle schools and find that school ranks vary substantially by scoring system. In this study, we build on this research by examining the same issue across a wider range of scoring systems and in a broader array of empirical contexts in the United States and globally.

2.2 Value-Added Models (VAMs) and Sensitivity to Model Specification

Consider a standard VAM:

$$\text{post}_{jk} = \beta_1 \text{pre}_{jk} + u_k + e_{jk} \tag{2}$$

$$e_{jk} \sim N(0, \sigma_e^2). \tag{3}$$

Here, post_{jk} is the endline test score for student j in teacher or school k, pre_{jk} is the baseline test score, β_1 is the strength of the relationship between baseline and endline scores, and e_{jk} is the residual error. u_k is the covariate-adjusted effect of school or teacher k on year end test scores, thus representing the "value added" to the learning process. While many approaches to VAMs are possible, such as gain scores, random effects, jackknife procedures, or the use of multiple pretests (Chetty, Friedman, & Rockoff, 2014a; Koedel et al., 2015), we maintain focus on Equation 2 for clarity and to focus on the impacts of test scoring method, holding all other factors fixed, as our arguments do not depend on the VAM specification used.

The identification of u_k as a causal effect of cluster k on student test scores depends on the extent to which the covariates capture relevant pre-existing differences between students in their sorting to clusters that also affect academic achievement. In other words, the identification strategy underlying a VAM is a selection on observables framework (Bacher-Hicks & Koedel, 2023; Castellano & Ho, 2015; Rothstein, 2009). The extent to which selection on observables is a realistic assumption is contested, though some evidence using experimental and quasi-experimental designs suggests that

VAMs can capture causal effects under certain conditions (Angrist et al., 2024; Chetty, Friedman, & Rockoff, 2014a, 2014b; Kane et al., 2013). Here, we focus not on the causal identification of VAMs, but the sensitivity of VAMs to model specification, as even under random assignment of students to clusters, model specification may still play an important role in the inferences drawn from a VAM.

Various reviews have emphasized the sensitivity of VAMs to model specification (Koedel et al., 2015; Levy et al., 2019). For example, VAMs are sensitive to included covariates, functional form specifications, and the timing of tests within the school year (Atteberry & Mangan, 2020; Cunningham, 2014; Levy et al., 2023; Papay, 2011). While application of VAMs to item-level outcome data is relatively rare, researchers have used item-level data to examine the effects of rapid guessing, heterogeneous VA effects across test items, and latent variable approaches on VAMs (Gilbert, Himmelsbach, Miratrix, et al., 2025; Hawley et al., 2017; Jensen et al., 2018). In this study, we use item-level data to explore the extent to which VAM results are sensitive to differences in the manner in which the test is scored.

2.3 The VAM Approach to Accounting for Measurement Error and Model Uncertainty

Traditional VAMs such as Equation 2 account for the uncertainty of the VA estimate by providing its standard error (SE), which is a function of the number of students per teacher or school and the residual variance σ_e^2. The SE from a single-timepoint model such as Equation 2 provides the expected sampling variability of the VA estimate, had different students been tested (Gilbert, Himmelsbach, Miratrix, et al., 2025).

Here, we focus on a complementary source of variation, often denoted "model uncertainty," a term typically associated with Bayesian Model Averaging (BMA, Clyde and George, 2004; Wasserman, 2000). That is, constructing a test score from item response data implies a statistical model (whether implicit or explicit), and the validity of SE is contingent upon the model being appropriately specified. For example, IRT models for dichotomous items include 1PL, 2PL, and 3PL approaches, which differ in whether the items vary in only their difficulty, or also in their discriminating power and the presence of lower asymptotes. IRT models for polytomous items include even more possibilities such as the rating scale, partial credit, and graded response models. However, the true model is generally unknown. Therefore, any VA estimate combines two classes of uncertainty: the traditional sampling variation of how the VA estimate would differ had alternative students been tested, and the model uncertainty of how the VA estimate would differ had an alternative scoring approach been used.

A simple method to adjust the traditional SE for model uncertainty is to calculate the variance of scores produced by different model and add this variation to the SE under the simplifying assumption that both sources of variation are independent. That is, we first estimate the variance of VA estimates for each unit k, where s indexes the VA estimate derived from scoring model s and there are S total scores (e.g., 1PL, 2PL, 3PL, etc.):

$$\sigma_k^2 = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{u}_{ks} - \overline{\hat{u}_k})^2. \tag{4}$$

We then define an "adjusted" SE for unit k by adding σ_k^2 to the variance of the VA estimate:

$$\text{Adjusted SE}_k = \sqrt{\text{SE}_k^2 + \sigma_k^2}. \tag{5}$$

In essence, Equation 5 is akin to a Generalizability Theory framework that treats the scoring model itself as a source of variation (Brennan, 1992, 2001).

We note that σ_k^2 and SE_k^2 are unlikely to be fully independent. The only thing that differs between VA estimates is the scoring model (e.g., 1PL vs. 2PL), so both the traditional SE and σ_k^2 are derived from the same underlying information. Any deviation caused by, say, item discrimination being estimated in a 2PL but not in a 1PL will partially overlap with what is captured by the SE. Thus, we view Equation 5 as a useful summary of how the magnitudes of the traditional SE and cross-model variation compare rather than a traditional additive variance decomposition. However, we run supplemental simulations where we generate scores from multiple IRT models fit to an unknown DGP and find that Equation 5 provides a latent trait coverage rate of 94.7%, in contrast to the typical

SEs, which provide a slightly lower coverage rate of 94.2%, suggesting that our approximation works reasonably to first approximation.

Our proposed approach is conceptually similar to that of Rights et al. (2018), who adjust the SEs of person scores derived from IRT models to account for model uncertainty. When VA estimates across scoring model are identical, $\sigma_k^2 = 0$ and the adjusted SE is equal to the traditional SE. As variation due to scoring model increases, the adjusted SE inflates accordingly to account the added model uncertainty. While the approach of Equation 5 is both simple to implement and transparent, it is somewhat ad hoc compared to BMA approaches that weight each model by its fit to the data. We return to this issue in our discussion.

2.4 Research Questions

We examine two primary research questions through an application to 18 education datasets with baseline and endline item responses:

1. How variable are VA estimates fit to models from different scores in a wide range of empirical data?
2. What dataset features predict this variation?

3 Methods

3.1 Data Sources

We draw our datasets from the Item Response Warehouse (IRW), a public repository of item-level datasets (Domingue, Braginsky, et al., 2025). We survey 18 datasets from education RCTs with common academic subjects (e.g., reading, math, science, etc.) as outcomes (Gilbert, Himmelsbach, Miratrix, et al., 2025; Gilbert, Himmelsbach, Soland, et al., 2025). All datasets contain dichotomous item responses at baseline and endline time points, and students are clustered in a higher level unit such as classrooms or schools. To maximize comparability across studies and isolate the effect of scoring method on the results, we limit the samples to students and items represented at both timepoints and remove any zero-variance items.

We summarize the datasets in Table 1. The data include 1,678,170 posttest item responses from 85,985 respondents (some of whom are represented more than once when studies contain more than one outcome measure). The datasets span a range of age groups, outcome measures, and geographic locations. While the original RCTs have some program evaluation aim, here, we use these data to examine the impact of scoring decisions on the properties of VA estimates.

Table 1: Summary of Empirical Datasets

Dataset	N	K	I	$\frac{N}{K}$	Cluster	Location	Age	Outcome
13: Romero et al. (2020)	3381	178	12	18.99	school	Liberia	Elementary	Literacy
14: Romero et al. (2020)	3381	178	40	18.99	school	Liberia	Elementary	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	8	21.34	class	USA	G5	Math
23: Bang et al. (2023)	886	41	31	21.61	class	USA	K-G1	Math
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
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
76: Thai et al. (2022)	428	20	11	21.4	class	USA	K	Math
78: Cabell et al. (2025)	1075	47	168	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	33	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	17	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 our references (Domingue, Braginsky, et al., 2025; Gilbert, Himmelsbach, Miratrix, et al., 2025; Gilbert, Himmelsbach, Soland, et al., 2025). We include the original dataset IDs in our tables and figures to facilitate replicability and comparability with the source studies.

3.2 VAM Scoring Approaches

To examine the effects of different scoring decisions on VAM estimates, we score item responses in several ways. These decisions mirror the steps in Section 2. We detail how we vary each scoring approach step by step.

3.2.1 Step 1. Whether to Use an Explicit Measurement Model

In the first step/decision point, we produce scores by taking the mean of the item responses. We then compare VAM results using mean scores to those using IRT-based measurement models, detailed below.

3.2.2 Step 2. Which Measurement Model to Use

Generally, we fit two broad types of measurement models using IRT. The first is a unidimensional model, fit separately to pre and post timepoints. We parameterized that model in several ways as well, including using 1PL, 2PL, and 3PL models. The second type is a longitudinal multidimensional IRT model (we will refer to it hereafter as a LIRT model). This model is detailed extensively in Kuhfeld and Soland (2022), therefore we do not provide much technical detail here. The model includes latent variables for each timepoint. Longitudinal measurement invariance is imposed such that the item parameters for a given item are consistent across timepoints. For example, we use a 2PL with item difficulty and discrimination parameters constrained to equality over time. Using these constraints allows us to fix the mean and variance of the latent variable at the first timepoint to zero and one respectively, then freely estimate the means and variances of the latent variables at later timepoints. These means and variances are estimated during calibration then treated as priors when scoring using EAP approaches.

3.2.3 Step 3. How to Estimate Model Parameters

Across studies, sample sizes are large enough to fit IRT models and we use Bayesian estimation to calibrate item parameters. We include a log-normal(0, 1) prior on the a_i parameter to ensure that all discrimination parameters are positive and to shrink extreme a_i parameters toward 1 and a normal(0, 2) prior on the b_i parameters. We fit 1PL, 2PL, and 3PL models to the data.

3.2.4 Step 4. How to Produce Scores

For all IRT models, we produce scores in three ways. First, we use MLE scoring. When scores are undefined because an examinee got all of the items right or wrong, we use the common practice of replacing undefined values with an arbitrary minimum/maximum (here, + or - 3 SDs). Second, we use Expected a Posteriori (EAP) scores because they minimize total error and, unlike maximum likelihood scores, are defined when a student answers all items correctly or incorrectly (Bock & Mislevy, 1982). Third, we use test characteristic curve (TCC) scores, based on the EAP scores. For TCC scores, we round to the nearest integer to mimic common practices in state accountability systems (e.g., Cognia and Massachusetts Department of Elementary and Secondary Education, 2024). We re-standardize all scores at pre and post to mean 0 variance 1 in the observed sample in a constrained Bayes approach (Ghosh, 1992) to reduce the impact of differential shrinkage and to maximize comparability of effect sizes across scoring models. We use the `mirt` package in R to fit the IRT models and extract the scores (Chalmers, 2012). We summarize the application of these four steps in our empirical analyses in Table 2. Note that the elements of Step 2 are only applied when using IRT scoring and that Step 4 is partially crossed with the IRT models from Step 2 (MLE and EAP for all models, and TCC based on the 2PL only). Thus, we obtain 10 VA scores for each cluster in our analysis.

Table 2: Scoring Decisions Applied to Our Empirical Datasets

Step	Empirical Application
1. Whether to Use an Explicit Measurement Model	Mean Scores vs. IRT Scores
2. Which Measurement Model to Use	1PL, 2PL, 3PL, LIRT (2PL)
3. How to Estimate Model Parameters	Bayesian
4. How to Produce Scores	EAP, MLE, TCC (2PL EAP)

3.3 VAM Estimation

We fit Equation 2 to each score from each dataset using the `lm` function in R and extract the VA estimates and their associated SEs for further analysis. Across models, we use the same scoring method for both baseline and endline, both to limit the number of comparisons and because it is unlikely that in practice evaluators would use different scoring systems across time points. Our analysis is not intended to be an exhaustive exploration of all possible scoring methods but rather illustrative of practical differences between common scoring methods across a wide range of empirical datasets. We then calculate the following metrics from the VA estimates from each dataset and scoring model for further analysis:

1. **Range and consistency of VA percentile ranks**. We calculate the VA percentile rank separately for each dataset and scoring model and calculate the range for each unit. For example, if school A has a minimum percentile rank of 20 and a maximum percentile rank of 30 across the scoring methods, the range for school A is 10 percentile points. We then calculate the intraclass correlation (ICC) of the percentile ranks to obtain a measure of VA rank stability in each dataset using the following equation:

$$\text{rank}_{ks} = \beta_0 + u_k + e_{ks} \tag{6}$$

$$u_k \sim N(0, \sigma_u^2) \tag{7}$$

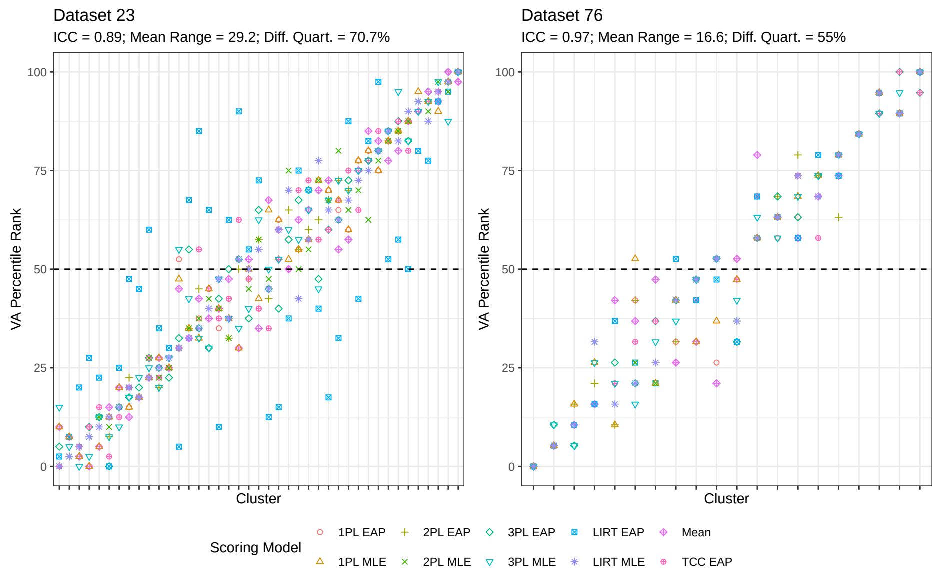
$$e_{ks} \sim N(0, \sigma_e^2), \tag{8}$$

   where rank_{ks} is the VA percentile rank of cluster k based on score s, u_k is the deviation of cluster k from the grand mean β_0, and e_{ks} is a residual. The ICC is defined as $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$.

2. **Range and consistency of VA estimates**. We apply the same method as we use for the percentile ranks to the VA estimates themselves, in standard deviation units.

3. **Classification in multiple quartiles**. We assign each percentile rank to the corresponding quartile and create an indicator variable for whether a unit is ever ranked across multiple quartiles.

Figure 1: VA Percentile Ranks by Scoring Method for Two Empirical Datasets



Notes: The figure shows VA percentile ranks for all schools in datasets 23 and 76. The x-axis shows the school, sorted by mean percentile rank, and the y-axis shows the VA percentile rank. The different colors and shapes represent the different scoring models. ICC = intraclass correlation, Mean Range = mean range of percentile ranks, Diff. Quart. = proportion of units classified multiple quartiles.

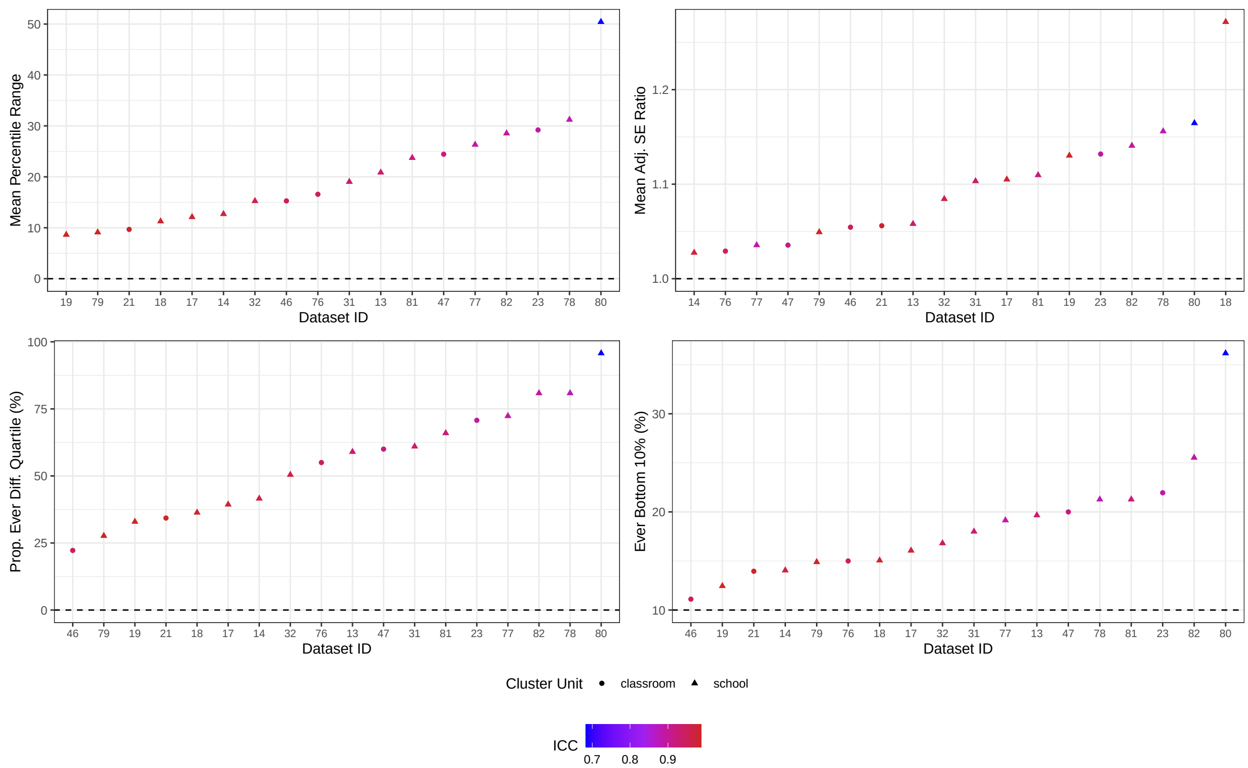
4. **Adjusted SE**. We apply Equation 5 to create adjusted SEs for each unit to determine how much the SEs inflate when adjusting for model uncertainty.

4 Results

Figure 1 contrasts VA percentile ranks by school two illustrative datasets, dataset 23 (Bang et al., 2023) and dataset 76 (Thai et al., 2022), both measuring math achievement in early elementary students. The percentile ranks for dataset 76 are more consistent than dataset 23 (ICC = .97 vs. .89), but nonetheless show significant variation, with an average range of VA percentile ranks of 16.6 points and 55% of classrooms classified in multiple quartiles. The percentile ranks for dataset 23 are much more variable, particularly for the LIRT EAP scores. We include identical scatterplots for all datasets in our supplement.

Figure 2 shows summary results for the 18 empirical datasets. The top left panel shows the mean VA percentile rank range, the top right panel shows the mean inflation of the adjusted SE over the

Figure 2: Effects of Scoring Method on VA Stability across 18 Empirical Datasets



Notes: The x-axis shows the dataset ID, sorted according to the y-axis variable. The points are color-coded by the ICC on the VA percentile rank (mean = .93, range = [.68, .99]) and the shapes represent whether the cluster unit is a classroom or school.

conventional SE (as a ratio), the bottom left shows the percent of units classified in multiple quartiles, and the bottom right shows the percent of units ever in the bottom 10% of the VA distribution. While the results vary widely across datasets, the overall pattern is clear: VA ranks and categorizations are in many cases highly sensitive to the scoring method used. For example, in many datasets, over half of the units are categorized in multiple quartiles, mean percentile rank ranges exceed 20 points, and the accounting for model uncertainty across the scoring methods adds more than 10% to the conventional SEs. Given that all four metrics examined in Figure 2 relate to the stability of VA rankings, it is not surprising that datasets tend to be ranked similarly across plots. For example, dataset 80 (Cabell et al., 2025) shows the most variation in three out of the four metrics, whereas dataset 46 (Glatz et al., 2023) shows the lowest variation in two out of the four metrics.

We conclude with a metaregression model to determine what dataset features predict dispersion in VA results. We regress the mean VA percentile range for each dataset on person sample size, item sample size (both logged), data sparsity (the proportion of possible item responses answered by participants, where 0 indicates that all respondents answer all items), range of standardized regression coefficients on the baseline scores, and marginal reliability of the posttest derived from a 2PL model. Given the small sample size of 18 datasets, we examine each predictor separately. We find that only data sparsity significantly predicts dispersion. In particular, a positive one-percentage point difference in data sparsity predicts a +.39 point difference in VA percentile ranks ($p < .01$). Thus, differences between scoring methods are less extreme with complete item response data. We emphasize that these results should be interpreted cautiously given the relatively small sample size and the convenience nature of the sample.

5 Discussion

While VAMs are widely used and debated in education policy and accountability research, the practical consequences of test scoring decisions on VAM interpretation has remained relatively unexplored. In this study, we find wide dispersion in VA estimates due to differences in scoring methods. Many of our 18 empirical datasets show relatively low stability of VA ranks across scoring methods, with average percentile rank ranges of 20 percentage points and over 50% of clusters classified in more than one performance quartile, holding constant both students and item responses. In general, the use of concurrently calibrated longitudinal IRT (LIRT) EAP scores emerged as the most significantly different approach than the alternative, separately calibrated scores. Because the concurrent calibration disattenuates the correlation between scores across time points for measurement error, the resulting pre/post correlations were in general substantially higher for the LIRT scores and the VA estimates therefore function somewhat differently compared to the other scoring systems (Lockwood & McCaffrey, 2014).

Our findings align with critiques of VAMs as arbitrary instruments of accountability due to their often low reliability (Amrein-Beardsley, 2014; Yeh, 2012). Our finding that VA rankings can vary by around 20 percentile points due to seemingly idiosyncratic statistical choices alone suggests the need for caution in interpreting point estimates from a single VAM. However, we emphasize that our focus on scoring methods differs from classical conceptions of measurement error in and reliability of VA estimates. Traditional measurement error assumes that the results would have been different had students (or items, depending on the scenario) varied across replications (Gilbert, Himmelsbach, Miratrix, et al., 2025). Here, we hold both students and items constant and only vary the scoring method and find that slight differences in the underlying scoring model can yield a meaningfully different inferences about a teacher or school's relative effectiveness, an important finding with implications for accountability policy that may be based in part on VAMs (Rights et al., 2018). Our findings are also aligned with those of McNeish and colleagues, who show that even when correlations between scoring methods are near perfect, substantive inferences can nevertheless be affected (McNeish, 2022, 2024; McNeish & Wolf, 2020). Thus, scoring decisions form an important but as of yet underappreciated component of validity evidence for uses of individual VA estimates in accountability policy.

How should practitioners respond in the face of such substantial scoring variation? The traditional VA SEs are only valid conditional on the model used to generate the scores being correct. In general, however, the true model is unknown (Rights et al., 2018). We propose a simple adjustment to account for model uncertainty, namely, inflating the SE to account for variation across scoring methods using Equation 5. In effect, this ad hoc approach is similar in spirit to Bayesian Model Averaging (Rights et al., 2018; Wasserman, 2000), under which model results are weighted according to some fit criteria such as the BIC, or plausible values scoring, where each scoring model is analogous to one plausible value (Huang, 2024). Equation 5 is similar, except that it assigns equal weights to the models. Beyond the advantages of simplicity and transparency, some scoring models, such as the TCC or mean score, do not produce standard fit indices, and the LIRT model fit indices are not comparable to the other IRT models due to the concurrent calibration across both time points. Thus, simplicity and transparency in scoring may be beneficial to justify VAMs for use in public accountability systems (see Sijtsma et al., 2024a; Sijtsma et al., 2024b for similar comments on interpretability of sum vs. 2PL scores). Furthermore, even when traditional fit indices show strong evidence for one model over another (which is extremely common when sample sizes are large), gains in out-of-sample predictive performance may be negligible. For example, Domingue et al. (2024) show that the 3PL model provides essentially no benefit in out-of-sample predictive performance over the 2PL model, *even when the 3PL is the data-generating process*. Thus, considerations of the tradeoffs between model fit, interpretability, transparency, and simplicity should be weighed carefully depending on the use case (Bonifay & Cai, 2017; Gilbert, 2025a, 2025b).

An alternative approach would be to simply select the most flexible or complex model. For example, if the data are generated from a 1PL model, fitting a 2PL model will return equal item discriminations within the bounds of sampling error and essentially reduce to the 1PL. We note two issues with this approach. First, even in very large samples, some parameters, such as the pseudo-guessing parameter in the 3PL model, are still very difficult to estimate, particularly for easy items. Second, there is essentially no upper limit to model complexity: 4PL models (including upper asymptotes, Liao et al., 2012), 5PL models (allowing for asymmetry, Johnson and Verkuilen, 2024), and non-parametric approaches are also available (Sijtsma, 1998). Furthermore, many datasets are simply too small to estimate such models precisely. Similarly, while the LIRT scores are perhaps the most theoretically defensible by linking the scoring scales across time and disattenuating pre/post correlations for measurement error, these benefits come at the cost of the added assumption that the item parameters are invariant over time.

The wide array of analysis options means that the adjusted SEs will themselves vary depending on the researcher's choice of scoring approaches to compare. Therefore, it may be necessary for analysts to define a universe of plausible modeling approaches to keep estimation tractable, as in Table 2. For example, as described above, while 4PL and 5PL models are available (Johnson & Verkuilen, 2024; Liao et al., 2012), to our knowledge, they are not used in any US state testing system. While a full review of state testing practices is outside the scope of this study, we highlight three illustrative examples for context. Massachusetts uses TCC scoring based on a 2PL (Cognia & Massachusetts Department of Elementary and Secondary Education, 2024, p. 65), the Measures of Academic Progress (MAP)—used in many states—uses 1PL scoring based on a Bayesian procedure (NWEA, 2019), and the 2022 New York Regents ELA exam uses TCC scoring based on a 1PL (Pearson, 2022). Thus, for purposes of school accountability in the US, the use of 1PL, 2PL, and 3PL models with EAP and TCC scoring may be sufficient, and researchers conducting scoring sensitivity analyses should consider the intended context and use of the VAM to identify the set of models to be compared.

We note several limitations of our study. First, our data sources are somewhat different from settings in which VAMs are commonly applied, such as US state accountability systems. While the global representation is a strength from a generalizability perspective, the extent to which our results would replicate on longitudinal data from US states is an open question, though results from Ng and Koretz (2015) across two scoring methods suggest similar patterns in at least one US state testing context. Second, while we explored a range of scoring methods for the dichotomous item responses, our design was not exhaustive. For example, plausible values IRT scores may perform differently than the EAP and MLE scores examined here, but they still suffer from model uncertainty. Third, estimating variability in VA scores across IRT models requires item-level data. In practical settings with proprietary tests, such data may not be readily available to analysts, and therefore the issues identified in this study may not be directly addressable in many real world contexts. Fourth, some research has emphasized the importance of controlling for multiple pretests in VAMs (Lockwood & McCaffrey, 2007), and we use only single pretests in our analysis (i.e., the lagged version of the posttest, which is necessary for the LIRT models). The extent to which multiple pretests might interact with the issues presented in this study are worthy of further investigation. Last, to compute the adjusted SE, the analyst must generate scores and fit the relevant VAMs multiple times, which may be very tedious and computationally intensive in practice, though such concerns may be attenuated with moderate sample sizes and modern computational capacities. For example, on the first author's personal computer, running the 10 VAMs across 18 datasets took approximately 3 minutes. However, such concerns may become more salient in large-scale contexts with hundreds of thousands or millions of students.

In conclusion, VAMs continue to play an important role in education evaluation, assessment, and accountability research, but considerations of how test scoring models may impact VA inferences have remained relatively unexplored. Our results show that quantitative and qualitative judgments about the efficacy of schools and teachers can be highly dependent on the scoring model used. As such, appropriate use of VAMs requires careful attention to scoring issues and may demand larger standard errors that account for model uncertainty in addition to traditional measurement error.

References

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

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education* (1st ed.). Routledge. https://doi.org/10.4324/9780203409909

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

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

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

Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC. https://eric.ed.gov/?id=ED458219

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383.

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.

Bauer, D., & Curran, P. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications* (pp. 3–38).

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

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

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/cbo9780511490026

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

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

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4*(4), 384–414. https://doi.org/10.1162/edfp.2009.4.4.384

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

Camilli, G. (2018). IRT scoring and test blueprint fidelity. *Applied Psychological Measurement*, *42*(5), 393–400. https://doi.org/10.1177/0146621618754897

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. https://doi.org/10.3102/1076998614548485

Chalmers, R. P. (2012). **mirt** : A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

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

Clyde, M., & George, E. I. (2004). Model Uncertainty. *Statistical Science*, *19*(1), 81–94. https://doi.org/10.1214/088342304000000035

Cognia & Massachusetts Department of Elementary and Secondary Education. (2024). *2023 Next-Generation MCAS and MCAS-Alt Technical Report* (tech. rep.). Massachusetts Department of Elementary and Secondary Education. https://www.doe.mass.edu/mcas/tech/2023-nextgen-tech-report.pdf

Cuhadar, I. (2022). Sample size requirements for parameter recovery in the 4-parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives*, *20*(2), 57–72. https://doi.org/10.1080/15366367.2021.1934805

Cunningham, P. L. (2014). *The effects of value-added modeling decisions on estimates of teacher effectiveness* [Doctoral Dissertation]. University of Iowa. https://doi.org/10.17077/etd.w4zo69mi

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.

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L. A., Gilbert, J., Kanopka, K., Kapoor, R., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2025). An introduction to the Item Response Warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*. https://doi.org/10.31234/osf.io/7bd54

Domingue, B. W., Kanopka, K., Kapoor, R., Pohl, S., Chalmers, R. P., Rahal, C., & Rhemtulla, M. (2024). The InterModel Vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items. *Psychometrika*, *89*(3), 1034–1054. https://doi.org/10.1007/s11336-024-09977-2

Domingue, B. W., Kanopka, K., Ulitzsch, E., & Zhang, L. (2025). Implied probabilities of polytomous response functions for model-based prediction and comparison. *Behaviormetrika*, *52*, 683–705. https://doi.org/10.1007/s41237-025-00262-9

Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental evidence on four policies to increase learning at scale. *The Economic Journal*, ueae003. https://doi.org/10.1093/ej/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*.

Edwards, K. D., & Soland, J. (2024). How scoring approaches impact estimates of growth in the presence of survey item ceiling effects. *Applied Psychological Measurement*, *48*(3), 147–164. https://doi.org/10.1177/01466216241238749

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, *87*(418), 533–540. https://doi.org/10.1080/01621459.1992.10475236

Gilbert, J. B. (2025a). Estimating treatment effects with the explanatory item response model. *Journal of Research on Educational Effectiveness*, *18*(1), 166–184. https://doi.org/10.1080/19345747.2023.2287601

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

Gilbert, J. B., Himmelsbach, Z., Miratrix, L. W., Ho, A. D., & Domingue, B. W. (2025). Item-level heterogeneity in value added models: Implications for reliability, cross-study comparability, and effect sizes [edworkingpapers.com]. https://doi.org/10.26300/EZ4Q-FS31

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

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.

Gorter, R., Fox, J.-P., Riet, G. T., Heymans, M., & Twisk, J. (2020). Latent growth modeling of IRT versus CTT measured longitudinal latent variables. *Statistical Methods in Medical Research*, *29*(4), 962–986. https://doi.org/10.1177/0962280219856375

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

Huang, F. L. (2024). Using plausible values when fitting multilevel models with large-scale assessment data using R. *Large-scale Assessments in Education*, *12*(1), 7. https://doi.org/10.1186/s40536-024-00192-0

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

Jessen, A., Ho, A. D., Corrales, C. E., Yueh, B., & Shin, J. J. (2018). Improving measurement efficiency of the Inner EAR scale with item response theory. *Otolaryngology–Head and Neck Surgery*, *158*(6), 1093–1100.

Johnson, P. J., & Verkuilen, J. (2024). Fisher information-based item difficulty and discrimination indices for binary item response models. In H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative Psychology* (pp. 177–188, Vol. 452). Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-55548-0_17

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

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

Kuhfeld, M., & Soland, J. (2022). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*, *27*(2), 234–260. https://doi.org/10.1037/met0000367

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

Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, *40*(10), 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Lockwood, J., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, *1*(none). https://doi.org/10.1214/07-EJS057

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. https://doi.org/10.3102/1076998613509405

Markus, K. A., & Borsboom, D. (2024). *Frontiers of test validity theory: Measurement, causation, and meaning* (2nd ed.). Routledge. https://doi.org/10.4324/9781003398219

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62. https://doi.org/10.3102/10769986031001035

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.

McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, *55*(8), 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D. (2024). Practical implications of sum scores being psychometrics' greatest accomplishment. *Psychometrika*, *89*(4), 1148–1169. https://doi.org/10.1007/s11336-024-09988-z

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

Nalbandyan, R., Gilbert, J. B., Franco, V. R., & Domingue, B. W. (2024). Signposts on the Path from Nominal to Ordinal Scales. https://doi.org/10.31234/osf.io/zbv8f

Ng, H. L., & Koretz, D. (2015). Sensitivity of School-Performance Ratings to Scaling Decisions. *Applied Measurement in Education*, *28*(4), 330–349. https://doi.org/10.1080/08957347.2015.1062764

NWEA. (2019). *MAP Growth Technical Report* (tech. rep.). https://www.nwea.org/uploads/2021/11/MAP-Growth-Technical-Report-2019_NWEA.pdf

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

Pearson. (2022). *New York State Regents Examination in English Language Arts 2022 Technical Report* (tech. rep.). https://www.nysed.gov/sites/default/files/programs/state-assessment/english-language-arts-technical-report-2022.pdf

Rhemtulla, M., & Savalei, V. (2025). Estimated factor scores are not true factor scores. *Multivariate Behavioral Research*, 1–22. https://doi.org/10.1080/00273171.2024.2444943

Rights, J. D., Sterba, S. K., Cho, S.-J., & Preacher, K. J. (2018). Addressing model uncertainty in item response theory person scores through model averaging. *Behaviormetrika*, *45*(2), 495–503. https://doi.org/10.1007/s41237-018-0052-1

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

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024a). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika*, *89*(1), 84–117. https://doi.org/https://doi.org/10.1007/s11336-024-09964-7

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*(1), 3–31. https://doi.org/10.1177/014662169802200101

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024b). Rejoinder to McNeish and Mislevy: What does psychological measurement require? *Psychometrika*, *89*(4), 1175–1185. https://doi.org/10.1007/s11336-024-10004-7

Soland, J. (2017). Is teacher value added a matter of scale? The practical consequences of treating an ordinal scale as interval for estimation of teacher effects. *Applied Measurement in Education*, *30*(1), 52–70. https://doi.org/10.1080/08957347.2016.1247844

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, *45*(5), 346–360.

Soland, J. (2022). Evidence that selecting an appropriate item response theory–based approach to scoring surveys can help avoid biased treatment effect estimates. *Educational and Psychological Measurement*, *82*(2), 376–403. https://doi.org/10.1177/00131644211007551

Soland, J. (2024). Item response theory models for difference-in-difference estimates (and whether they are worth the trouble). *Journal of Research on Educational Effectiveness*, *17*(2), 391–421. https://doi.org/10.1080/19345747.2023.2195413

Soland, J., Cole, V., Tavares, S., & Zhang, Q. (2025). Evidence that growth mixture model results are highly sensitive to scoring decisions. *Multivariate Behavioral Research*, *0*(0), 1–22. Retrieved April 30, 2025, from https://doi.org/10.1080/00273171.2024.2444955

Soland, J., Edwards, K., & Talbert, E. (2025). When should evaluators lose sleep over measurement? Toward establishing best practices. *Journal of Research on Educational Effectiveness*, *18*(3), 474–506. https://doi.org/10.1080/19345747.2024.2344011

Soland, J., Kuhfeld, M., & Edwards, K. (2024). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, *29*(5), 1003–1024.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

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