EdWorkingPaper No. 23-766

How Measurement Affects Causal Inference: Attenuation Bias is (Usually) More Important Than Scoring Weights

Joshua B. Gilbert
Harvard University

When analyzing treatment effects on test scores, researchers face many choices and competing guidance for scoring tests and modeling results. This study examines the impact of scoring choices through simulation and an empirical application. Results show that estimates from multiple methods applied to the same data will vary because two-step models using sum or factor scores provide attenuated standardized treatment effects compared to latent variable models. This bias dominates any other differences between models or features of the data generating process, such as the use of scoring weights. An errors-in-variables (EIV) correction removes the bias from two-step models. An empirical application to data from a randomized controlled trial demonstrates the sensitivity of the results to model selection. This study shows that the psychometric principles most consequential in causal inference are related to attenuation bias rather than optimal scoring weights.

VERSION: February 2024

Suggested citation: Gilbert, Joshua B.. (2024). How Measurement Affects Causal Inference: Attenuation Bias is (Usually) More Important Than Scoring Weights. (EdWorkingPaper: 23-766). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/4hah-6s55

How Measurement Affects Causal Inference: Attenuation Bias is (Usually) More Important Than Scoring Weights

Joshua B. Gilbert[1]

[1]Harvard Graduate School of Education

Abstract

When analyzing treatment effects on test scores, researchers face many choices and competing guidance for scoring tests and modeling results. This study examines the impact of scoring choices through simulation and an empirical application. Results show that estimates from multiple methods applied to the same data will vary because two-step models using sum or factor scores provide attenuated standardized treatment effects compared to latent variable models. This bias dominates any other differences between models or features of the data generating process, such as the use of scoring weights. An errors-in-variables (EIV) correction removes the bias from two-step models. An empirical application to data from a randomized controlled trial demonstrates the sensitivity of the results to model selection. This study shows that the psychometric principles most consequential in causal inference are related to attenuation bias rather than optimal scoring weights.

Keywords: causal inference, latent variable models, factor analysis, psychometrics, educational measurement

1 Introduction

When research results are sensitive to the choice of statistical model, they become dependent on researcher discretion, and bias can be introduced (King & Nielsen, 2019; Simmons et al., 2011; Wicherts et al., 2016). Researcher discretion is a particular challenge in educational research on test score outcomes because of the many approaches to scoring tests and modeling test score data, such as Classical Test Theory (CTT), Item Response Theory (IRT), Factor Analysis (FA), or Latent Variable Models (LVMs). Researcher-designed assessments in particular demand many decision points in the analysis process, raising the question of how sensitive results are to model selection and scoring decisions, especially in causal studies investigating intervention impacts on test score outcomes that aim to provide policy-relevant findings.

For a given causal research question, alternative statistical methods may provide defensible options for analysis, and varying results are expected. For instance, when modeling a binary outcome, logistic regression and the linear probability model may produce different results due to the contrasting assumptions of each model (Timoneda, 2021). Similarly, in the context of multisite randomized trials or meta-analyses, fixed effects and random effects estimators will produce different estimates of treatment effects due to the different estimands targeted by each model (Chan & Hedges, 2022; Miratrix et al., 2021; Skrondal & Rabe-Hesketh, 2004). While such differences in "estimates, estimators, and estimands" (Miratrix et al., 2021) are well understood in causal inference generally, the use of test score data as outcome measures demands additional consideration because scores are typically not of interest in themselves but rather as proxies for unobserved latent variables like mathematics or reading proficiency. Thus, researchers are faced with navigating a range of options for causal analysis of test score data and the challenge of interpreting differing results from models that theoretically target the same treatment effect on the latent trait. Furthermore, it is unclear whether some approaches are consistently superior to others or the tradeoffs of model selection depend on the circumstances (Gilbert, 2024; Hontangas et al., 2015).

As an example, consider the options for scoring an educational test to estimate a treatment effect on the latent trait represented by the test score. Both CTT sum scores and IRT- or FA-based scores use item responses to estimate a latent trait score for each student, which is then used in subsequent analysis. IRT- or FA-based methods, such as the two-parameter logistic (2PL) model or the congeneric factor model, theoretically provide more fine-grained distinctions among students by weighting item responses based on the information provided by the item, in contrast to sum scores, which treat different sets of correct answers as identical (Camilli, 2018). An analyst may alternatively apply LVM techniques, such as Structural Equation Modeling (SEM; Kline, 2023; Muthén, 2002) or Explanatory Item Response Modeling (EIRM; Briggs, 2008; Gilbert, 2024), to estimate measurement and regression models in a single step. Because all test scoring methods and LVMs target the same treatment effect on the latent trait, a key question is the extent to which theoretical differences between these models matter in causal analysis of test score outcome data. Correlations between IRT- and FA-based scores and CTT scores are typically above 0.90 (Soland et al., 2022, p. 8), which raises the question of whether the theoretical benefits of IRT- or FA-based scoring methods or LVMs are worth the added complexity, computational power, and interpretational challenges they may pose. Furthermore, no clear guidelines exist on which model researchers should prefer, particularly when the results conflict.

To illustrate the challenge facing the applied researcher, consider two recent publications on the implications of using sum scores versus factor scores as outcomes. On one side, McNeish and Wolf, 2020 argue that sum scores can have "adverse effects on validity, reliability, and qualitative classification" compared to FA-based scores. In contrast, Widaman and Revelle, 2023 argued that so long as the scale is unidimensional, sum scores "often have a solid psychometric basis and therefore are frequently quite adequate for psychological research". What is the applied researcher to do?

The purpose of this study is to provide both a concise and accessible review of the conceptual issues at play and practical guidance for researchers by evaluating the consequences of measurement modeling decisions on causal inference by determining which decision points in measurement modeling are most salient for analytic results. Results show that the issue of attenuation bias dominates the issue of scoring weights, and simpler models can perform better even under extreme circumstances. Our results align with recent studies showing that the marginal gains to more complex statistical models can be low and may not justify their increased complexity (e.g., Domingue, Kanopka, Kapoor, et al., 2022 in IRT; Castellano and Ho, 2015, in value-added modeling), and serve as a contrast with other work emphasizing the sensitivity of analytic results to measurement modeling choices in the analysis of psychometric data (McNeish & Wolf, 2020; Soland et al., 2022).

1.1 Classical Approaches to Measurement Error in Education Research

Measurement error is a widely studied phenomenon in education, with work the reliability of psychological tests going back many decades (Asher, 1974; Borsboom, 2005; Briggs, 2021; Cronbach, 1951; Lord & Novick, 1968), and has well-known consequences in statistical analysis (Fuller & Hidiroglou, 1978; Hutcheon et al., 2010; Liu, 1988). In the case of simple linear regression with two variables, error in independent (X, predictor) variables serves to attenuate regression coefficients toward 0, whereas error in dependent (Y, outcome) variables will not bias estimated regression coefficients, but will decrease precision and reduce statistical power by increasing residual variance, though these general rules of thumb do not always hold in more complex circumstances (Kline, 2023). Measurement error can be addressed using both classical and modern methods. For example, Errors-in-Variables (EIV) regression models (Gillard, 2010) use estimates of reliability to deattenuate the coefficients of predictor variables, and LVMs (Muthén, 2002) adjust for measurement error by simultaneously estimating the latent variable(s) and the regression model. While both EIV and LVM methods can correct for measurement error, some studies have shown that the LVM approach can provide more robust estimates of uncertainty than EIV methods (Gilbert, 2024; Lockwood & McCaffrey, 2020).

Measurement error in the dependent variable is sometimes ignored because it does not bias coefficients, but LVMs can also be applied to outcome variables and have been demonstrated to provide modest benefits to statistical power and more robust estimates of uncertainty than alternative approaches (Christensen, 2006; Rabbitt, 2018; Zwinderman, 1991), though benefits are context dependent (Gilbert, 2024). However, coefficients *are* downwardly biased when the dependent variable is standardized. Attenuation due to standardization is a particular issue in education research because test scores have no natural scale, and standardization allows for estimates of treatment effect size that can in principle be compared across studies or pooled in meta-analyses (Borenstein et al., 2021) and are often argued to be more interpretable than unstandardized coefficients (Schielzeth, 2010).

Standardization of the dependent variable Y attenuates regression coefficients because measurement error causes overdispersion in the standard deviation of Y, σ_Y. That is, σ_Y will be greater than the SD of the true latent trait scores σ_T because σ_Y contains the variation of σ_T plus measurement error σ_E, as summarized in the CTT variance decomposition $\sigma_Y^2 = \sigma_T^2 + \sigma_E^2$. We can precisely estimate the overdispersion of σ_Y with the CTT reliability formula, which defines reliability ρ as the ratio of true score variance (σ_T^2) to observed score variance (σ_Y^2): $\rho = \frac{\sigma_T^2}{\sigma_Y^2}$. Solving for σ_Y shows that $\sigma_Y = \frac{\sigma_T}{\sqrt{\rho}}$. Therefore, when we standardize an outcome variable such as a test score by dividing by its SD σ_Y, this value is too large by a factor of $\frac{1}{\sqrt{\rho}}$. Consequently, when measurement error in the outcome is present, standardized regression coefficients will be driven downward, and this bias can be corrected by dividing by $\sqrt{\rho}$. Applying this EIV correction deattenuates the standardized regression coefficient to what it would be if the test were perfectly reliable or of infinite length. This fact is not a new insight (Hedges, 1981), but it is nonetheless commonly ignored, or reserved for technical discussions (Borenstein et al., 2021). For example, in its section on reliability, the What

Works Clearinghouse Standards Handbook lists minimum thresholds for various reliability metrics (e.g., $\alpha \geq .50$ in Version 4.1 and $\alpha \geq .60$ in Version 5.0), but makes no mention of attenuation bias, in contrast to detailed explanation of the bias that arises from other sources, such as non-random attrition or baseline non-equivalence.[1] Crucially, attenuation bias is not solved by IRT or FA scoring procedures, because the shrinkage of the empirical Bayes estimation draws the distribution of estimated latent trait scores to the overall mean rather than the respective conditional (i.e., group-specific) means (Briggs, 2008; Soland, 2022). This problem is less severe but still present when using maximum likelihood scoring (Soland et al., 2022, p. 11). The solution is to apply an EIV correction by dividing the coefficients by $\sqrt{\rho}$, where ρ can be estimated as the internal consistency of the test (e.g., Cronbach's α) or to employ an LVM that directly adjusts for measurement error in the estimation procedure, as we will demonstrate.

1.2 Methods for Estimating Causal Effects on Test Score Outcome Data

1.2.1 Two-Step Procedures

In a two-step procedure, the latent trait of interest is estimated for each person and then analyzed as the outcome variable using a standard statistical model such as OLS regression (Christensen, 2006; Ye, 2016). For example, consider the following regression model, in which score_j represents an estimated latent trait score for person j (for persons $j = 1, ..., J$) and β_1 represents the average treatment effect (ATE):

$$\text{score}_j = \beta_0 + \beta_1 \text{treat}_j + \varepsilon_j \tag{1}$$

$$\varepsilon_j \sim N(0, \sigma_\varepsilon). \tag{2}$$

[1] Current and past WWC Standards Handbooks are available at the following URL: https://ies.ed.gov/ncee/wwc/handbooks.

score$_j$ may be generated in a CTT or IRT/FA framework. In CTT, a sum or mean score is used, such that the observed score across all items for items $i = 1, ..., I$ equals the sum of the responses $\sum_{i=1}^{I} \text{item}_i$ or the mean of the responses $\frac{1}{I}\sum_{i=1}^{I} \text{item}_i$. In IRT or FA, the latent trait estimate, denoted $\hat{\theta}$, is calculated by maximizing the likelihood of $\hat{\theta}$ given the estimated item parameters (i.e., item difficulty/factor intercept, item discrimination/factor loading, and pseudo-guessing) (Bock et al., 1997). Generally, the IRT scoring approach has been argued to be superior because IRT $\hat{\theta}$ estimates are on an interval rather than ordinal scale (Briggs, 2008; Ferrando & Chico, 2007; Harwell & Gatti, 2001; Jabrayilov et al., 2016; McNeish & Wolf, 2020) and IRT/FA models weight item responses by their discrimination parameters or factor loadings, thus maximizing the information in $\hat{\theta}$ (Camilli, 2018; Jessen et al., 2018). Empirically however, differences between CTT and IRT/FA scoring are often found to be minor (Sébille et al., 2010; Xu & Stone, 2012). One limitation of the two-step approach is that, regardless of what type of scoring procedure is used to estimate the latent trait, the outcome variable is treated as known when it contains error and therefore resulting regression coefficients will be biased when the outcome is standardized.

1.2.2 Simultaneous Estimation with Latent Variable Models (LVMs)

As an alternative to two-step procedures, LVMs enable the analyst to estimate measurement (psychometric) and regression (structural) models simultaneously and incorporate the effects of measurement error directly into the estimation procedure, for both predictors and outcomes (Kline, 2023; Muthén, 2002). For example, consider the following LVM for the analysis of a treatment effect on test score data,

$$Y_{ij} = \lambda_i(\theta_j + b_i) + \varepsilon_{ij} \tag{3}$$

$$\theta_j = \beta_0 + \beta_1 \text{treat}_j + \zeta_j \tag{4}$$

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon_i}) \tag{5}$$

$$\zeta_j \sim N(0, \sigma_\zeta) \tag{6}$$

in which the response Y to item i for person j is a function of latent person ability θ_j and item easiness parameters (factor intercepts) b_i, weighted by item discrimination parameters (factor loadings) λ_i and error term ε_{ij}. θ_j is in turn a function of the control group mean β_0, the ATE β_1, and unexplained or residual variance in person ability ζ_j. Thus, the ATE β_1 is estimated directly on the latent trait without the need to compute a summary score. When all $\lambda_i = 1$, and a logistic link function is employed, the LVM is equivalent to the One Parameter Logistic (1PL) Explanatory Item Response Model (EIRM; De Boeck and Wilson, 2016). When λ_i are freely estimated and an identity link function is used, the LVM is a linear Structural Equation Model (SEM). While LVMs such as the EIRM and SEM can be more complex to interpret than two-step approaches, LVMs automatically deattenuate estimates of standardized regression coefficients because, unlike σ_Y, σ_ζ is an unbiased estimator of the residual SD of θ_j, thus counteracting the effects of measurement error compared to regression on observed scores (Briggs, 2008; Christensen, 2006; Zwinderman, 1991), suggesting that LVMs may provide more accurate tests of between-group differences such as causal treatment effects.

In sum, the researcher faces many choices in model selection when test score data are used as outcomes in a causal inference context: to use a one-step or two-step approach, to weight or not to weight, to use CTT or IRT/FA, and so forth. While exploratory data analysis can shed light on, for example, whether a 1PL or 2PL IRT model is a better fit to the data, to what extent does allowing for varying item discriminations/loadings in the estimation of the latent trait score affect the bias, precision, and power of causal estimates? Are certain models consistently more robust than alternatives? This study seeks to shed light on these questions and leverage measurement principles for better application of causal inference in education research by using Monte Carlo simulation and an empirical application to examine the performance of several models under varying conditions of test score construction.

2 Methods

2.1 Data Generating process

The simulation and data analysis procedures were implemented in R. In total, we simulated 18,000 data sets (1,000 data sets per 18 data-generating conditions) and applied four analytic models—sum score, factor score, equal loading SEM, and variable loading SEM—to each, for a total of 90,000 results. We employed a full factorial design to assess the performance of each model under a range of treatment effect sizes and items of varying discriminating power. To maintain focus on the contrasts between the models and the effects of test characteristics, we fixed the number of subjects at 500 and the number of items at 10 to represent a moderate sample size and moderate test length. The latent trait scores θ_j were drawn from $N(0 + \beta_1 \text{treat}_j, 1)$ and the factor intercepts b_i were drawn from $N(0, 1)$. The latent variables were converted to continuous observed scores for each item using Equation 3. The residual SD for each item σ_{ε_i} was defined as $\sqrt{1 - \lambda_i^2}$. The simulation factors include null, moderate, and large treatment effect sizes (0, 0.2, or 0.4 SDs on the latent trait), moderate and high average factor loadings ($\mu_\lambda = 0.4, 0.6$), and constant, moderately variable, or highly variable factor loadings ($\lambda_i \sim \text{Unif}(\mu_\lambda - x, \mu_\lambda + x)$ where $x = 0, 0.3, 0.6$).

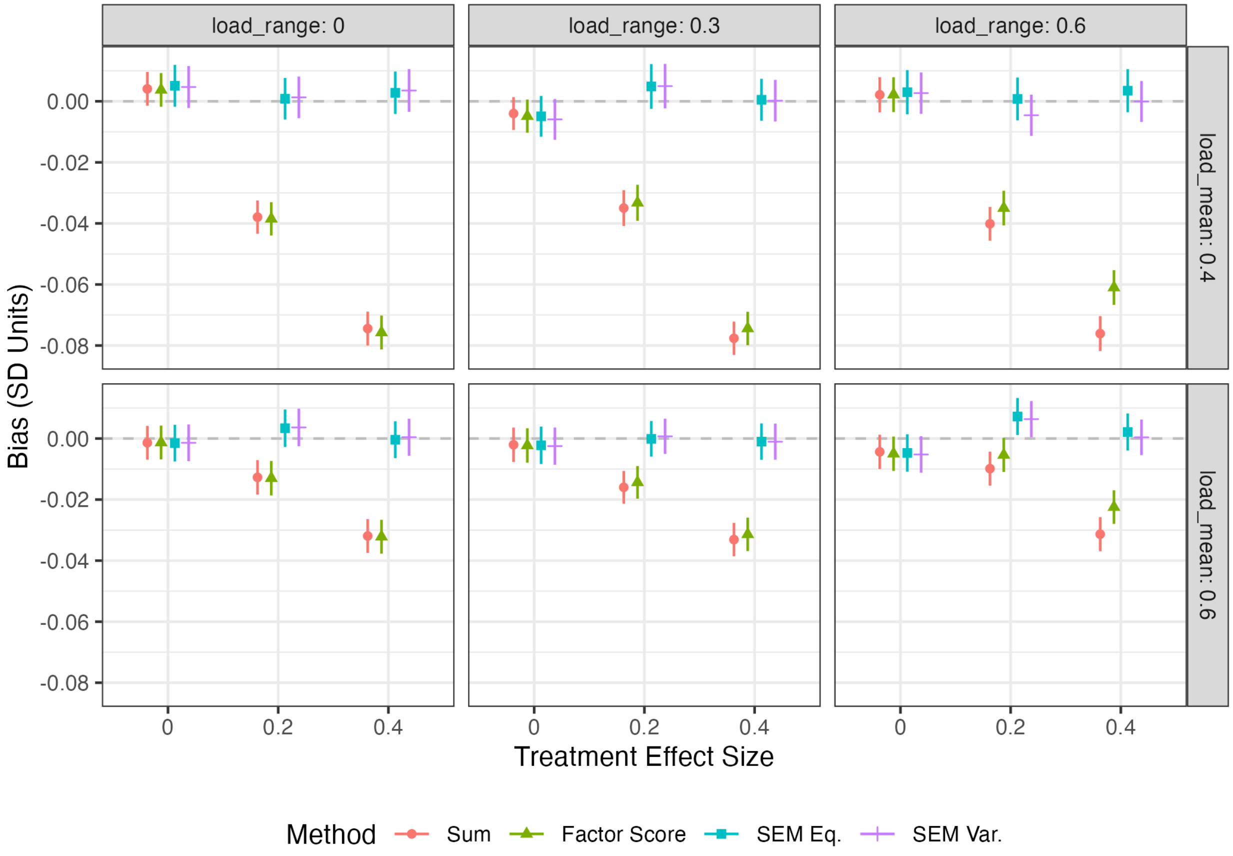
For each simulated data set, we estimated the treatment effect and associated z-statistic, p-value, and whether the null hypothesis was rejected under each model. The models for the sum score and factor scores are equivalent OLS regression models and the SEMs are estimated using maximum likelihood with fixed factor intercepts. In all models, the parameter of interest is the ATE β_1, and the errors are assumed to be normally distributed with mean 0 and constant variance and uncorrelated with the predictors. We also calculated α for each simulated test as an estimate of ρ to assess the effect of applying EIV corrections to the two-step models. To render each ATE comparable, we divide β_1 by the RMSE of the regression model to standardize the coefficients, as the RMSE represents the estimated pooled (i.e., within-group) standard deviation of the latent trait θ_j. Thus, the standardized coefficients are equivalent to Cohen's d effect size.

3 Results

Figure 1 shows the mean estimated bias and Monte Carlo 95% confidence intervals for each method across all simulation conditions. We see that when the ATE is 0, bias is negligible across all conditions. However, when the ATE is positive, the two-step procedures are downwardly biased, the bias is proportional to the treatment effect size, and the bias is more severe when the loadings are lower because lower average loadings translates to lower test reliability. In contrast, the LVMs do not show the same pattern of attenuation and are approximately unbiased across all conditions. Crucially, the performance of the SEM assuming equal factor loadings is indistinguishable from the SEM allowing for variable loadings, even when the range of loadings is high. The factor score allowing for variable weights only slightly outperforms the sum score when the range of loadings is highest, but its performance is nonetheless bested by the equal-weight SEM.

These results clearly illustrate that attenuation due to measurement error with standardized outcome variables is a more serious concern than the decision of whether to weight or not to weight the item responses. When we correct the two-step procedures for measurement error by dividing the coefficients by $\sqrt{\alpha}$ as shown in Figure 2, we find that the performance of the sum score is indistinguishable from the LVMs. Interestingly, the the EIV correction appears to overcorrect the factor score when the loadings are extremely variable. This occurs because

Figure 1: Estimated Bias by Method, Standardized Scores



the calculation of α assumes equal loadings and provides a lower bound on test reliability. When this assumption is not met, $\rho > \alpha$ so dividing by $\sqrt{\alpha}$ provides an overcorrection for the factor score model, though the bias is still small in absolute magnitude.

Figures 3, 4, 5, and 6 show the performance of the models in terms of absolute precision (the SD of the point estimates), standard error calibration (the mean model-based SE as a proportion of the true SE), false positive rates, and statistical power, with the EIV correction applied to the two-step models. We see that differences between all models are minimal according to these metrics, even when item loadings are highly variable, suggesting that once the attenuation bias of the two-step models has been corrected, the choice of model does not appear to have strong impacts on the other statistical properties of the ATE. Note that

Figure 2: Estimated Bias by Method, EIV Correction Applied to Two-Step Scores

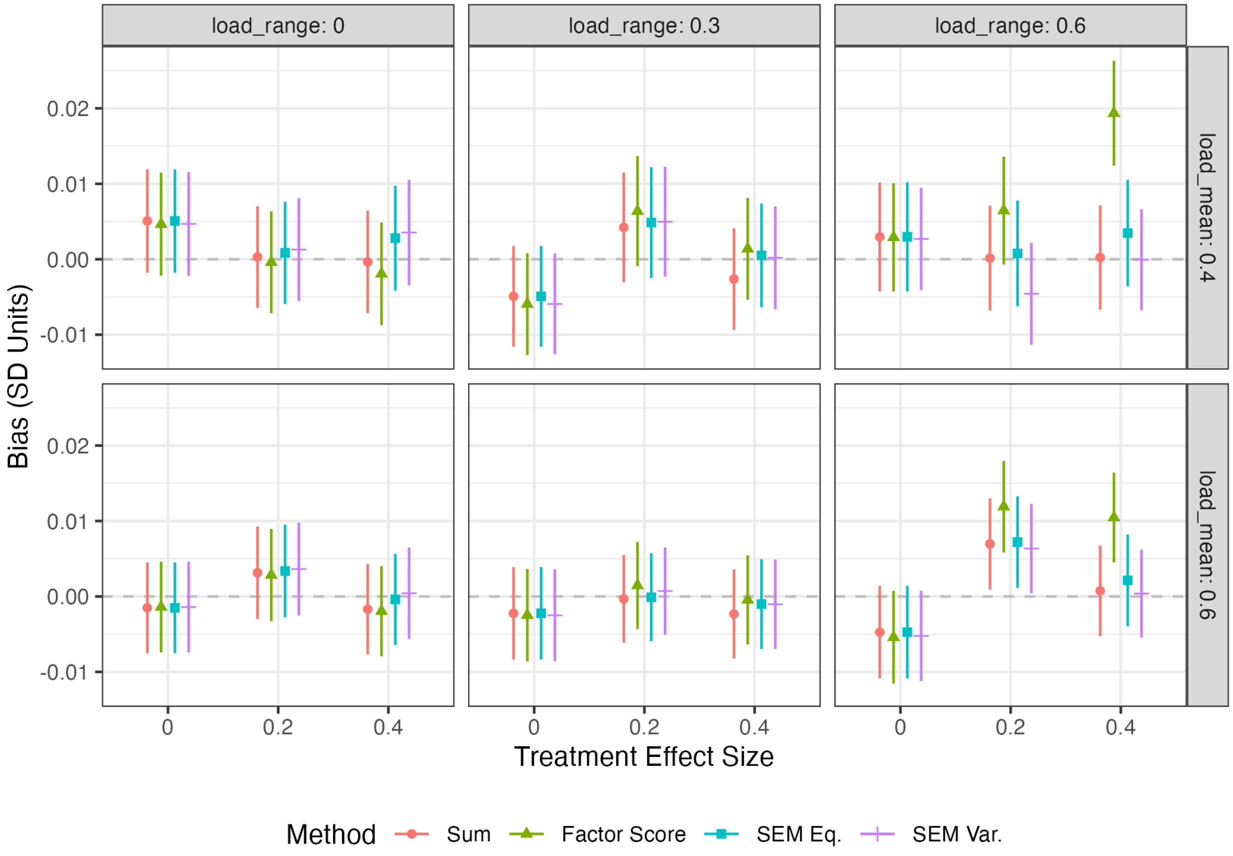
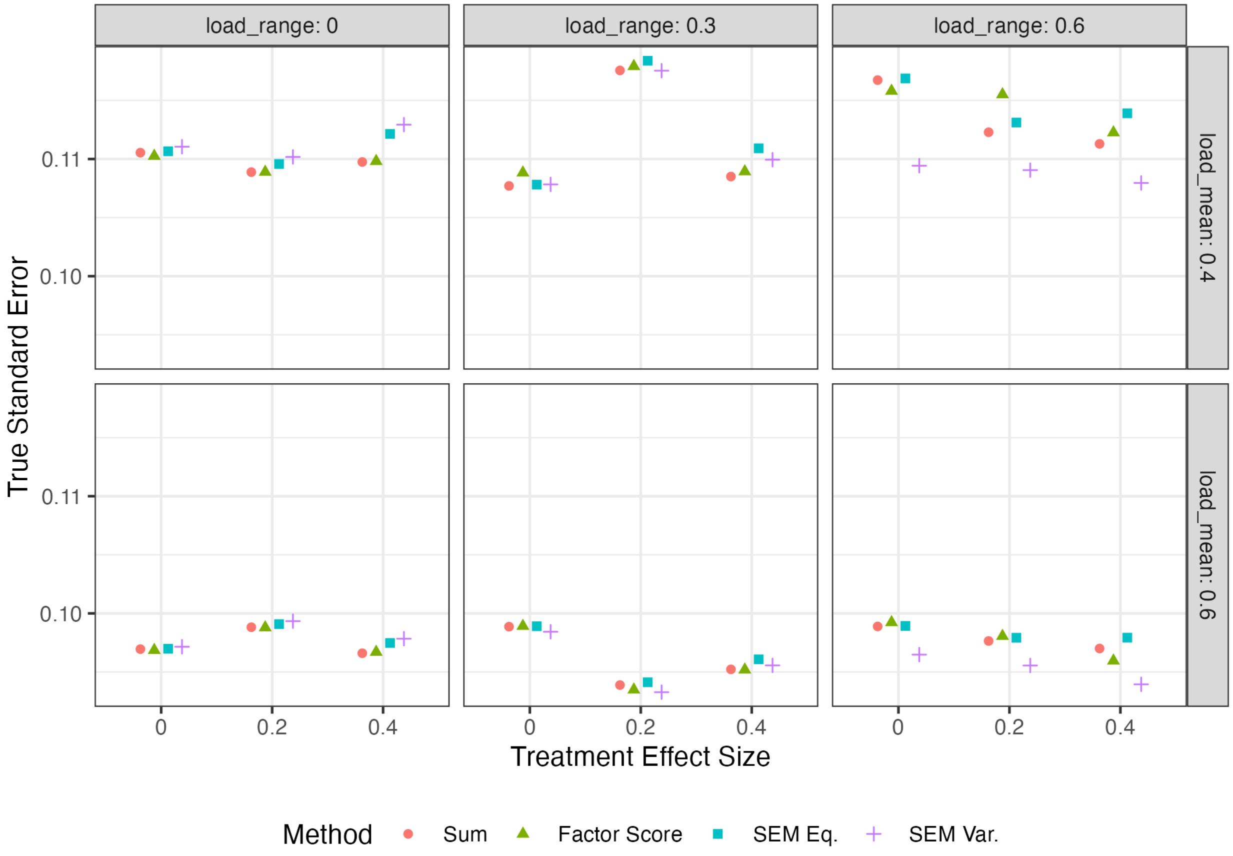


Figure 6 shows the results for the treatment effect size of 0.2 only because near ceiling levels of power were achieved at a treatment effect size of 0.4.

4 Empirical Application

To illustrate how the issues of model selection can play out in practice, we employ a public use file from Kim et al., 2023 that explores the causal effect of the Model of Reading Engagement (MORE) literacy intervention on 2$^{\text{nd}}$ grade students' test scores on a researcher-designed reading comprehension assessment. The assessment included three reading passages and 20 multiple choice items, and the study employed a cluster-randomized design with 30 schools and 2174 students. The authors assess the intention-to-treat (ITT) effect of the MORE

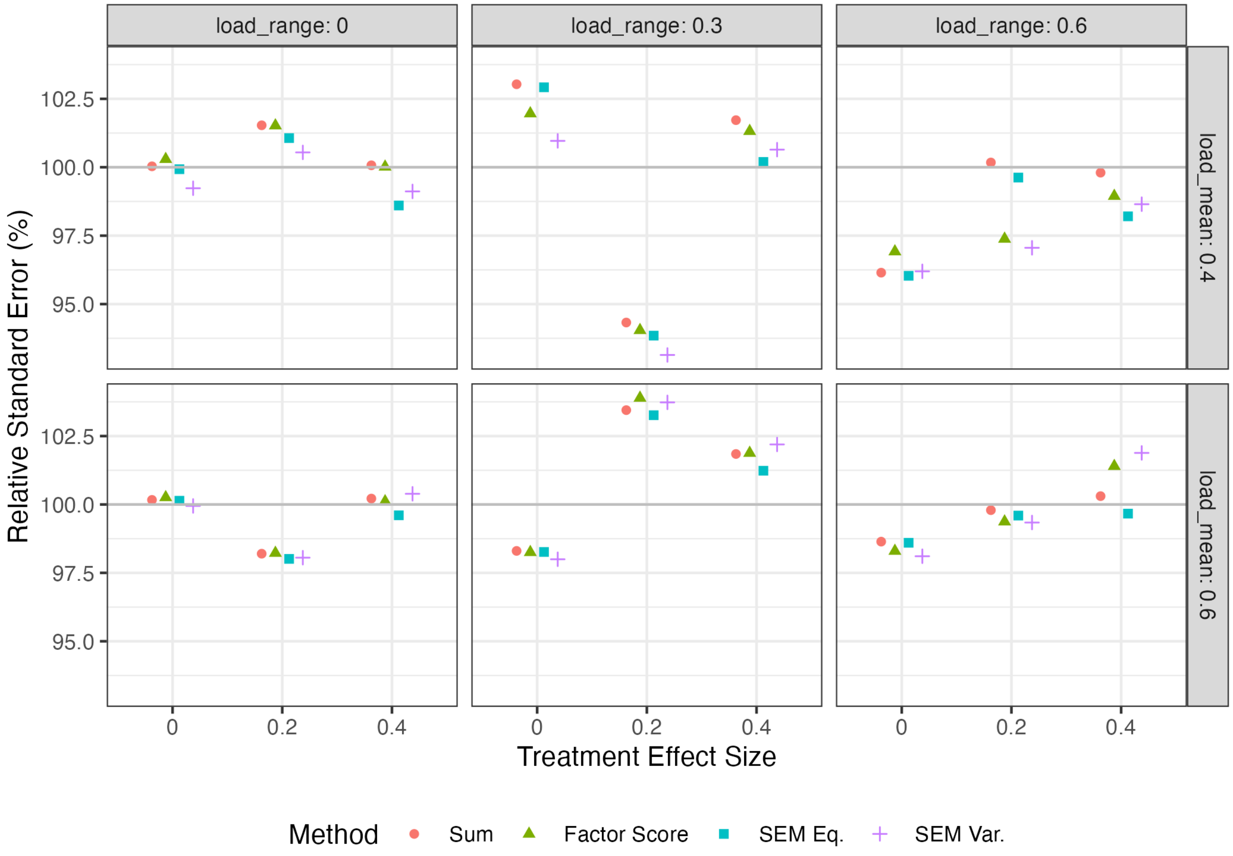
Figure 3: Estimated Standard Errors by Method



intervention by fitting a multilevel model in which the outcome was the standardized sum score, which they justify by noting that the correlation between 2PL IRT-based scores and sum scores was 0.98 (p. 14).

The MORE assessment data differs in three key respects from the setup of our simulations. First, as a cluster-randomized trial, the data have a hierarchical structure with treatment assigned at the school level. Second, the authors included a rich set of demographic covariates such as pretest scores, race, gender, SES, and other student characteristics to increase the precision of their ITT estimates. Third, the test items in this data set are dichotomous (0 = incorrect, 1 = correct) rather than continuous. In our reanalysis of the MORE study data, we simplify by ignoring the school-level clustering and the demographic control variables to maintain focus on the treatment effect point estimates derived from different models, as

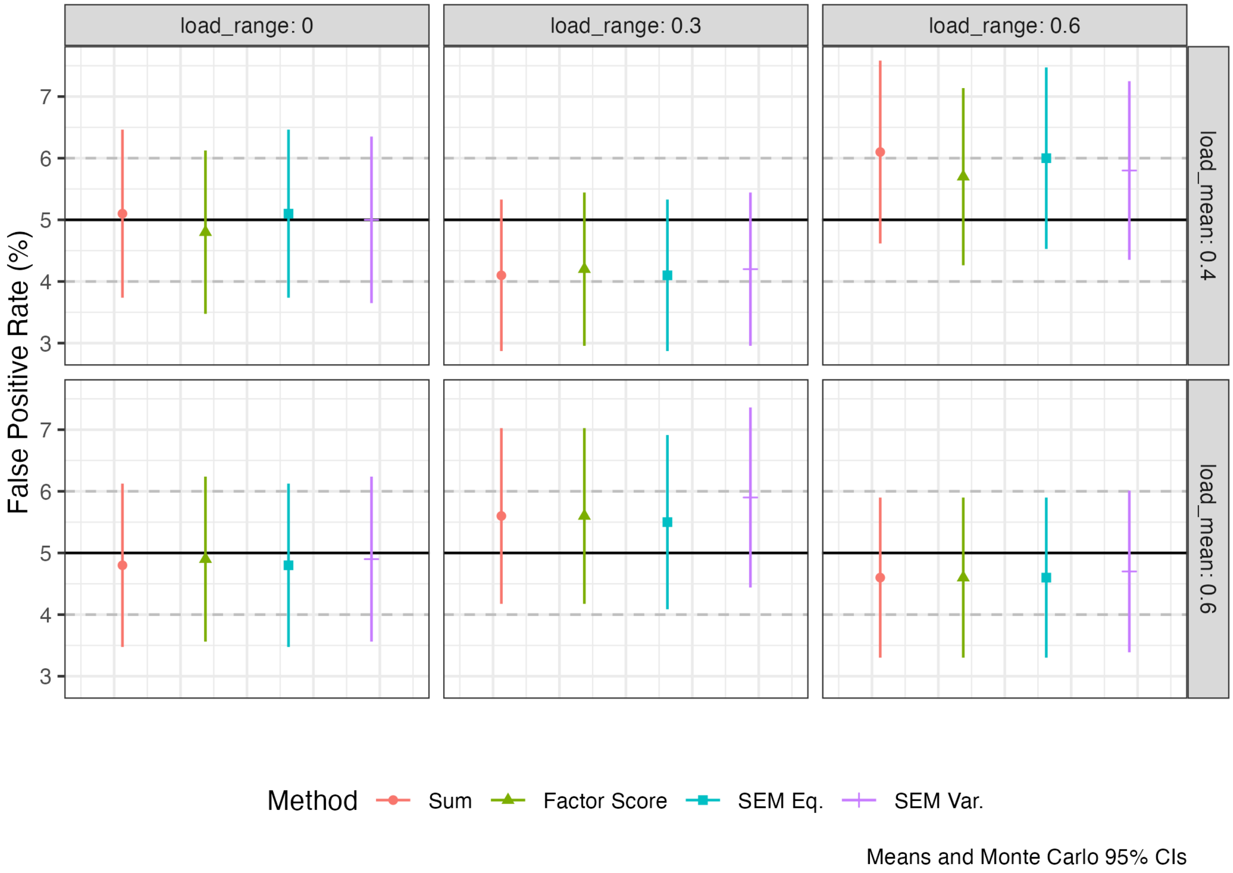
Figure 4: Estimated Standard Error Calibration by Method



this approach provides a closer analog to the simulation models. We maintain the pretest covariate to increase the precision of the estimates, and because the authors reported a slight baseline imbalance in pretest scores, despite the randomization. We also add a few scoring models relevant for dichotomous data by using 1PL and 2PL IRT models to generate the factor scores and allowing for a logistic link function in the equal loading LVM. We fit models of the following general form,

$$y_j = \beta_0 + \beta_1 \text{treat}_j + \beta_2 \text{pretest}_j + \varepsilon_j \tag{7}$$

Figure 5: Estimated False Positive Rates by Method



in which y_j is the sum or factor score for student j, β_1 is the treatment effect, β_2 is the coefficient for pretest reading score, and ε_j is the student-level residual. The LVMs are analogous but model the treatment effect on the latent trait directly using the 20 items as indicators for the latent trait.

While we cannot know the true value of the treatment effect on the latent trait in the empirical data, we can still examine the results of the different analytic models explored in the simulation and see how sensitive the results are to the modeling choice. As noted earlier, the authors use a standardized sum score as their outcome variable and do not make any adjustments for measurement error. The internal consistency of the 20-item reading comprehension assessment is estimated at 0.78 suggesting an attenuation bias of 13% ($\frac{1}{\sqrt{.78}} = 1.13$) for two-step models of standardized scores.

Figure 6: Estimated Statistical Power by Method

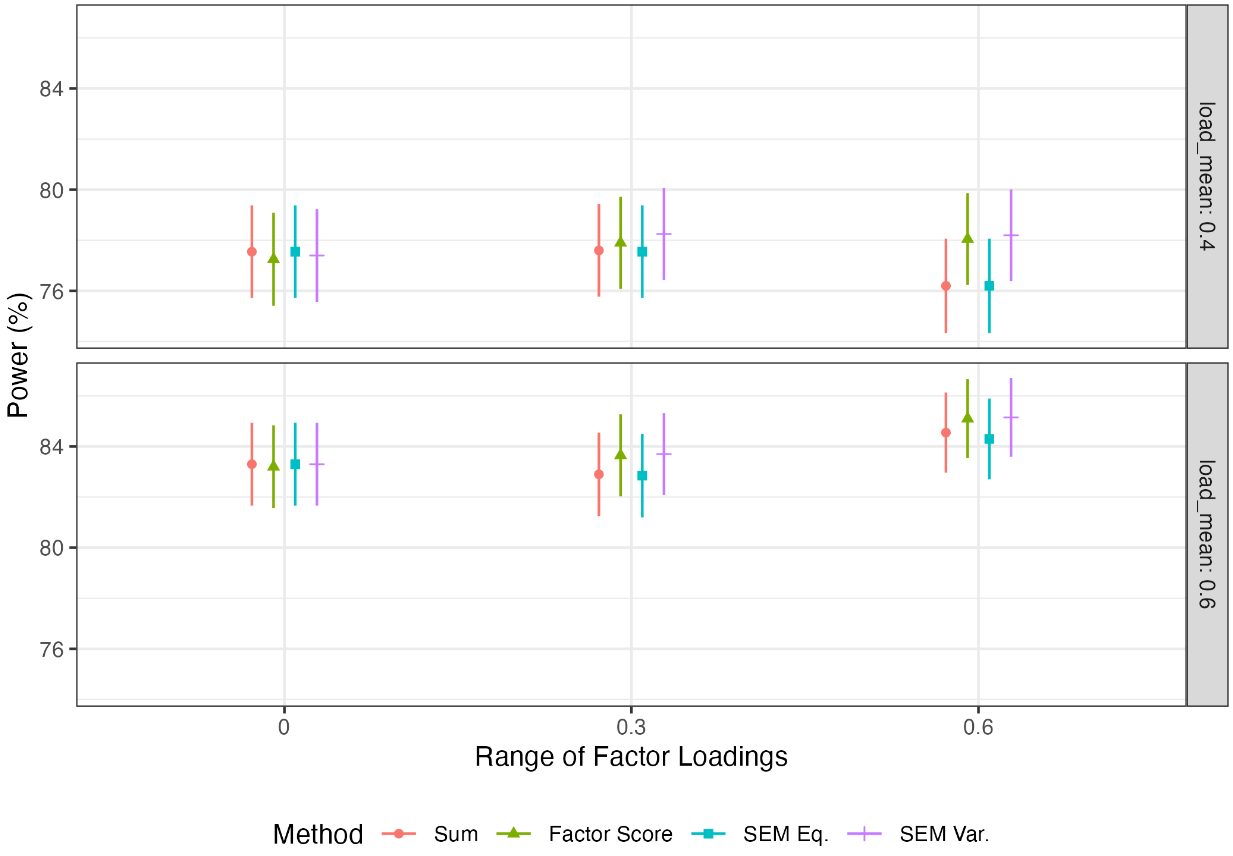
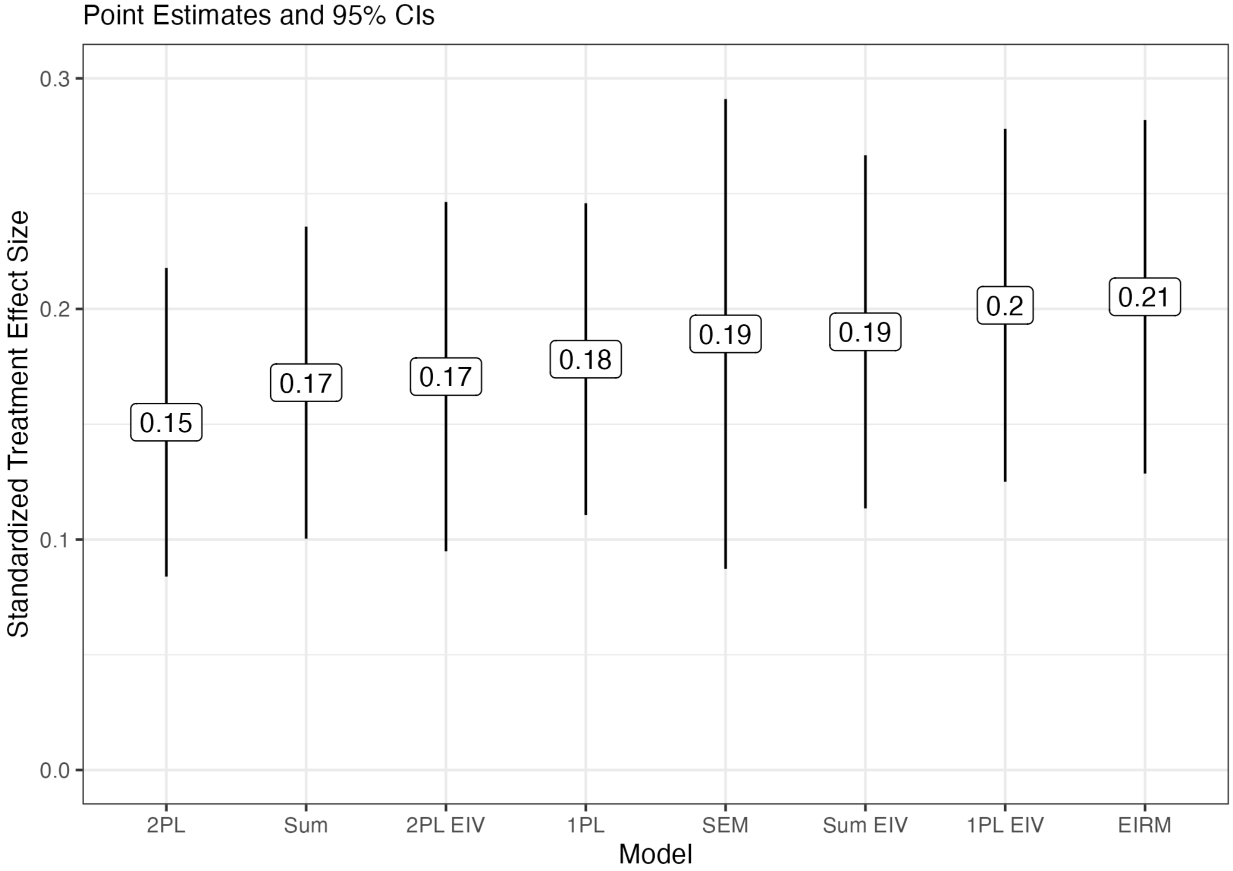


Figure 7 shows the point estimates and 95% CIs for the standardized treatment effect using the various models. Though CIs overlap considerably, and all reject the null hypothesis, clearly, measurement matters, as there is meaningful variation in the of point estimates, with a low end of 0.15SDs for the unadjusted 2PL IRT score model, vs. 0.21SDs for the EIRM, a non-trivial difference of 0.06SDs (+40%). Given the results above, the EIRM or the EIV-adjusted sum score are likely to be more accurate estimates of the true treatment effect on the latent trait.

Figure 7: Estimated Treatment Effects Derived from the MORE Data



5 Discussion

Because test scores are a noisy proxy of a latent trait of interest, they suffer from measurement error, which results in negatively biased treatment effect estimates when outcome variables are standardized. Simulation results show that when applied to test score data with different properties, the bias is substantial when treatment effect sizes are high. However, when the EIV correction is applied and the standardized coefficients are divided by $\sqrt{\alpha}$, differences in model performance are negligible under most conditions. Thus, the very process that makes varying statistical models comparable to one another—standardization—biases two-step models, and the effect of this bias dominates other features of the data generating process, including the variability of factor loadings used to create scoring weights.

Our interpretation of these results is that researchers may be overly focused on second-order measurement issues, such as the use of variable factor loadings that function as optimal scoring weights, rather than the first-order issue of attenuation of standardized coefficients for measurement error in the outcome variable. That is, when the EIV correction is applied, differences between the simplest standardized sum score model and the more complex LVMs are negligible in terms of bias, precision, and statistical power in the estimation of treatment effects, and this result holds even when the variability of factor loadings is high. Thus, when causal inference at a single time point is the primary goal, the use of sum scores with the EIV correction is likely to be sufficient for many applications.

These results should not be interpreted as evidence that IRT or LVMs have no purpose. Clearly, IRT/FA methods are essential for piloting measures, identifying poorly functioning items (Jessen et al., 2018), differential item functioning analysis (Osterlind & Everson, 2009), vertical scaling (Briggs & Domingue, 2013), linking (Lee & Lee, 2018), and addressing missing data (Gilbert, 2024), and LVMs can easily be expanded to incorporate complex relationships among many latent variables or multidimensional constructs at several time points (Kline, 2023). A particularly valuable use case for LVMs in causal inference would be settings in which treatment may impact individual items and the LVM can provide insights on treatment heterogeneity, such as "item-level heterogeneous treatment effects" that would be masked in a two-step analysis (Ahmed et al., 2023; Gilbert, 2023; Gilbert, Kim, & Miratrix, 2023b; Sales et al., 2021), differential growth by item type (Briggs, 2021; Gilbert, Kim, & Miratrix, 2023a; Naumann et al., 2014), or the appropriate interpretation of interaction effects (Domingue, Kanopka, Trejo, et al., 2022; Gilbert, Miratrix, et al., 2023). However, when all students receive the same items at a single time point, and only average treatment effects are of interest, the results appear relatively insensitive to the methods employed when the EIV correction is applied. Therefore, the benefits of interpretability and computational complexity may favor the EIV-corrected standardized sum score in many straightforward causal inference applications, a finding contrary to arguments that the sum score is a suboptimal choice because the constraint of equal factor loadings imposed by the sum score is rarely met in real data (McNeish & Wolf, 2020).

While the results of this study provide strong evidence for the importance of EIV corrections in two-step analyses of standardized test score outcome variables, several limitations merit consideration. First, the data generating process employed in this study examines the simple case of individual randomization with no covariates beyond the treatment indicator, and thus may be extended to explore how measurement model selection may impact the estimation of heterogeneous treatment effects, the effects of predictive covariates, multilevel structures such as multi-site or cluster-randomized trials, or alternative experimental and quasi-experimental contexts such as regression discontinuity, difference-in-differences, instrumental variables, and longitudinal analyses, though an emerging literature on the synthesis of latent variable and causal inference methods has begun to shed light on these areas (Gilbert, Kim, & Miratrix, 2023a; Gilbert, Miratrix, et al., 2023; Kuhfeld & Soland, 2022, 2023; Miratrix et al., 2021; Rabbitt, 2018; Soland, 2022, 2023; Soland et al., 2023).

In conclusion, results of causal analyses of test score data are sensitive to model selection, and the effects of attenuation bias are much more consequential than the use of scoring weights. When researchers do not adjust for measurement error with EIV corrections LVMs, standardized treatment effect estimates will be downwardly biased and thus understate estimates of treatment impact. When the EIV correction is applied, the impact of model selection will be reduced, demonstrating how the application of psychometric principles can improve causal inference in education research.

6 Data Availability

1. A replication toolkit is available via Research Box for researchers interested in replicating or extending the analyses in this study at the following URL: https://researchbox.org/2289&PEER_REVIEW_passcode=HAEVNQ.

2. The public use data from Kim et al., 2023 are available at the following URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LAWFFU.

3. The supplemental materials for Kim et al., 2023 are available at the following URL: https://supp.apa.org/psycarticles/supplemental/edu0000751/edu0000751_supp.html and include the assessment itself as well as psychometric analysis of the items, including CTT, IRT, and factor analysis.

References

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Benjamin, W. (2023). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials.

Asher, H. B. (1974). Some consequences of measurement error in survey data. *American Journal of Political Science*, 469–485.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). Irt estimation of domain scores. *Journal of educational measurement*, *34*(3), 197–211.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

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. (2021). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, *38*(6), 551–576.

Camilli, G. (2018). Irt scoring and test blueprint fidelity. *Applied Psychological Measurement*, *42*(5), 393–400.

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.

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

Christensen, K. B. (2006). From rasch scores to regression. *Journal of Applied Measurement*, *7*(2), 184.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

De Boeck, P., & Wilson, M. R. (2016). Explanatory response models. In *Handbook of item response theory* (pp. 593–608). Chapman; Hall/CRC.

Domingue, B. W., Kanopka, K., Kapoor, R., Pohl, S., Chalmers, P., Rahal, C., & Rhemtulla, M. (2022). The intermodel vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items.

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*.

Ferrando, P. J., & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between irt and ctt. *Psicologica: International Journal of Methodology and Experimental Psychology*, *28*(2), 237–257.

Fuller, W. A., & Hidiroglou, M. A. (1978). Regression estimation after correcting for attenuation. *Journal of the American statistical Association*, *73*(361), 99–104.

Gilbert, J. B. (2023). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in r. *Behavior Research Methods*.

Gilbert, J. B. (2024). Estimating treatment effects with the explanatory item response model. *Journal of Research on Educational Effectiveness*, *0*(0), 1–19. https://doi.org/10.1080/19345747.2023.2287601

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023a). Leveraging item parameter drift to assess transfer effects in vocabulary learning. (868). http://www.edworkingpapers.com/ai23-868

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023b). 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*, 10769986231171710.

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2023). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity.

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *REVSTAT-Statistical Journal*, *8*(1), 57–80.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *journal of Educational Statistics*, *6*(2), 107–128.

Hontangas, P. M., De La Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and irt scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *Bmj*, *340*.

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied psychological measurement*, *40*(8), 559–572.

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.

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.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, *27*(4), 435–454.

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

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.

Kuhfeld, M., & Soland, J. (2023). Scoring assessments in multisite randomized control trials: Examining the sensitivity of treatment effect estimates to measurement choices. *Psychological Methods*.

Lee, W.-C., & Lee, G. (2018). Irt linking and equating. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 639–673.

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology*, *127*(4), 864–874.

Lockwood, J., & McCaffrey, D. F. (2020). Recommendations about estimating errors-in-variables regression in stata. *The Stata Journal*, *20*(1), 116–130.

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

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

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*(1), 270–308.

Muthén, B. O. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, *29*, 81–117.

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.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.

Rabbitt, M. P. (2018). Causal inference with latent variables from the rasch model as outcomes. *Measurement*, *120*, 193–205.

Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). The effect of an intelligent tutor on performance on specific posttest problems. *International Educational Data Mining Society*.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113.

Sébille, V., Hardouin, J.-B., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory (ctt) and item response theory (irt)-based approaches for the comparison of patient-reported outcomes in two groups of patients-a simulation study. *BMC medical research methodology*, *10*(1), 1–10.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

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.

Soland, J. (2023). Item response theory models for difference-in-difference estimates (and whether they are worth the trouble). *Journal of Research on Educational Effectiveness*, 1–31.

Soland, J., Johnson, A., & Talbert, E. (2023). Regression discontinuity designs in a latent variable framework. *Psychological Methods*, *28*(3), 691.

Soland, J., Kuhfeld, M., & Edwards, K. (2022). How survey scoring decisions can influence your study's results: A trip through the irt looking glass. *Psychological Methods*.

Timoneda, J. C. (2021). Estimating group fixed effects in panel data with a binary dependent variable: How the lpm outperforms logistic regression in rare events data. *Social Science Research*, *93*, 102486.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, 1832.

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788–806.

Xu, T., & Stone, C. A. (2012). Using irt trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468.

Ye, F. (2016). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 43–61.

Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.