METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

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

Joshua B. Gilbert[1]

[1] *Harvard Graduate School of Education, Harvard University, Cambridge, MA, USA.*

Methodology, 2025, Vol. 21(2), 91–122, https://doi.org/10.5964/meth.15773

Received: 2024-10-05 • **Accepted:** 2025-05-06 • **Published (VoR):** 2025-06-30

Handling Editor: Juraj Medzihorsky, Durham University, Durham, United Kingdom

Corresponding Author: Joshua B. Gilbert, 13 Appian Way, Cambridge, MA 02138, USA. E-mail: joshua_gilbert@g.harvard.edu

Supplementary Materials: Data [see Index of Supplementary Materials]

Abstract

When analyzing treatment effects on outcome variables constructed from psychometric instruments (e.g., educational test scores, psychological surveys, or patient reported outcomes), researchers face many choices and competing guidance for scoring the measures and modeling results. This study examines the impact of outcome measure scoring and modeling approaches 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 10 datasets from randomized controlled trials 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 outcome scoring weights.

Keywords

causal inference, latent variable models, factor analysis, psychometrics, measurement

When research results are sensitive to the choice of statistical model, they become dependent on researcher discretion, and bias can be introduced (Gelman & Loken, 2013;

King & Nielsen, 2019; Simmons et al., 2011; Wicherts et al., 2016). Researcher discretion is a particular challenge in evaluation research on outcomes derived from psychometric instruments, such as educational tests, psychological surveys, and patient-reported outcomes, because of the many approaches to scoring the outcome measures and accounting for measurement error in the results, such as Classical Test Theory (CTT), Item Response Theory (IRT), Factor Analysis (FA), or Latent Variable Models (LVMs). Researcher-designed outcome measures in particular demand many decision points in the analysis process, raising the question of how sensitive results are to model selection and outcome scoring decisions, especially in causal studies investigating intervention impacts on outcomes that aim to provide policy-relevant findings. While reviews in several fields such as medicine, education, and organizational research show a relative lack of attention to issues of measurement in general and call for better measurement practices (Brakenhoff et al., 2018; Cox & Kelcey, 2019; Flake et al., 2017; Flake & Fried, 2020; Pedersen et al., 2025; Spybrook et al., 2016), the implications of measurement principles for causal inference, policy, and program evaluation are less prominent (Shear & Briggs, 2024; Soland, Kuhfeld, & Edwards, 2024; Soland, Edwards, & Talbert, 2024).

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 psychometric measures as outcome variables demands additional consideration because observed scores are typically not of interest in themselves but rather as proxies for unobserved latent variables such as academic achievement. Thus, researchers are faced with navigating a range of options for causal analysis of psychometric outcome 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 if the tradeoffs of model selection depend on the circumstances (Gilbert, 2024a; Hontangas et al., 2015).

As an example, consider the options for scoring an educational test to estimate a treatment effect on the latent trait of academic achievement imperfectly represented by the observed 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 (i.e., item discrimination or factor loading) provided by the item, in contrast to sum scores, which treat different sets of correct answers as identical (Camilli, 2018; Hambleton & Van der Linden, 1982; Lord, 1980; Lord & Novick, 1968; Thissen & Wainer, 2001). Alternatively, LVM techniques, such as Structural Equation Modeling (SEM; Kline, 2023; Muthén, 2002) or Explanatory Item Response Modeling (EIRM; Briggs, 2008; De Boeck, 2004; De Boeck & Wilson, 2016; Gilbert, 2024a; Wilson et al., 2008) estimate the 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 (Lu et al., 2005; Soland, Kuhfeld, & Edwards, 2024), 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 in statistical models. 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 because sum scores implicitly assume that each item contributes equally to the estimation of the latent trait, an assumption that is unlikely to be met in many empirical applications. In contrast, Widaman and Revelle (2023) argue 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”. Such competing claims, expanded in further publications (McNeish, 2022, 2023, 2024; Sijtsma et al., 2024), provide a challenge for the applied researcher working with outcome data derived from psychometric measures.

The purpose of this study is to provide both a concise and accessible review of the conceptual issues at play and practical guidance for evaluation researchers by exploring the consequences of outcome 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. In other words, our results suggest that accounting for measurement error in the outcome variable is a first-order concern in causal inference, in contrast to second-order issues of measurement “model error”, in which an incorrect measurement model is applied to generate scores for the outcome variable (Liu & Pek, 2024), such as using equally-weighted scores when other approaches are a better fit to the data. Our results align with 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., 2024 in IRT; Castellano & Ho, 2015, in value-added modeling; Widaman & Revelle, 2023 in psychological measurement), 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, Kuhfeld, & Edwards, 2024).

Classical Approaches to Measurement Error in Evaluation Research

Measurement error is a widely studied phenomenon, with work on the reliability of educational and psychological tests going back many decades (Asher, 1974; Bollen, 1989; 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 with both classical and modern methods. For example, Errors-in-Variables (EIV) regression models (Carroll et al., 2009; Gillard, 2010) use estimates of reliability to adjust 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, 2024a; Lockwood & McCaffrey, 2014).

Measurement error in the dependent variable is sometimes ignored because it does not bias coefficients (Cox & Kelcey, 2019), but LVMs can also be applied to outcome variables and can 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, 2024a). However, coefficients *are* downwardly biased when the dependent variable is standardized. Attenuation due to standardization is a particular issue in evaluation research because most test scores, psychological surveys, and patient reported outcomes 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., 2009) 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 $\sigma_Y^2 = \sigma_T^2 + \sigma_E^2$ (Brennan, 2010; DeVellis, 2006; Hambleton & Jones, 1993;

Jackson, 1973; Lewis, 2006; Traub, 1997). 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.

Attenuation due to standardization is not a new insight (Cole & Preacher, 2014; Hedges, 1981; Shear & Briggs, 2024), but it is nonetheless commonly ignored, or reserved for technical discussions (Borenstein et al., 2009) and comparatively less emphasized in practical guides for researchers. For example, in its section on reliability, the Institute of Education Sciences' (IES) *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 resulting scores still contain measurement error. The problem can be further compounded by expected a posteriori (EAP) scoring methods because shrinkage of the empirical Bayes estimation draws the distribution of estimated latent trait scores to the overall mean across treatment and control groups rather than the respective means of each group (Briggs, 2008; Soland, 2022). This problem is less severe but still present when using maximum likelihood (ML) scoring (Soland, Kuhfeld, & Edwards, 2024, p. 11), though ML scoring raises other issues such as undefined scores for respondents with "perfect" scores (i.e., all items answered correctly or incorrectly on an educational test). 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 ω) (Hedges, 1981; Shear & Briggs, 2024), or to employ an LVM that directly adjusts for measurement error in the estimation procedure, as we will demonstrate.

Methods for Estimating Causal Effects on Psychometric Outcome Data

Estimands and Estimators

Consider outcome θ_j for person j ($j = 1, \ldots, J$). Under the potential outcomes framework (Imbens & Rubin, 2015; Rubin, 2005), the individual causal effect of binary treat-

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

ment T_j on person j is $\tau_j \equiv \theta_j(1) - \theta_j(0)$, where 1 indicates the treatment counterfactual and 0 indicates the control counterfactual.[2] Because only one counterfactual is observed, τ_j is unobservable. The target estimand of causal analyses is therefore typically the average treatment effect (ATE), defined as $\overline{\tau} = \frac{1}{J}\Sigma_{j=1}^{J}(\theta_j(1) - \theta_j(0))$.

Random assignment of the treatment ensures that treatment status is independent of the potential outcomes. Therefore, we can estimate $\overline{\tau}$ as a difference in means between the treated and control groups. Practically, we can use a simple linear regression model as our estimator for $\overline{\tau}$, in which T_j is an indicator variable for the treatment status of person j, β_0 is the mean of the control group, β_1 is the difference in means between the groups, and ε_j is the error term (Angrist & Pischke, 2009; Imbens & Rubin, 2015; Murnane & Willett, 2010; Rosenbaum, 2017):

$$\theta_j = \beta_0 + \beta_1 T_j + \varepsilon_j\,. \tag{1}$$

When θ_j is observed, the difference in means approach provided by Equation 1 is standard. However, when θ_j represents an unobserved variable, such as mathematical ability, extroversion, or depression, Equation 1 is no longer estimable (Stoetzer et al., 2024). The two primary approaches to estimating causal effects on latent variables are two-step procedures and simultaneous estimation, to which we now turn.

Two-Step Procedures

In a two-step procedure, the latent trait of interest is first 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 and β_1 represents the average treatment effect (ATE):

$$\text{score}_j = \beta_0 + \beta_1 \text{treat}_j + \varepsilon_j \tag{2}$$

$$\varepsilon_j \sim N(0, \sigma_\varepsilon)\,. \tag{3}$$

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 $i = 1, \ldots, I$ 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 $\widehat{\theta}$, is calculated by maximizing the likelihood of $\widehat{\theta}$ given the estimated item parameters (Bock et al., 1997). Generally, the IRT scoring approach has been argued to be superior to CTT approaches because IRT $\widehat{\theta}$ estimates are on an interval rather than ordinal scale (Ferrando & Chico, 2007; Harwell & Gatti, 2001; Jabrayilov et

[2] We follow the notation of Gilbert, Himmelsbach, et al. (2025a).

al., 2016; McNeish & Wolf, 2020).[3] Furthermore, scores provided by IRT/FA models weight the contributions of item responses to $\hat{\theta}$ by their discrimination parameters or factor loadings, thus maximizing the information in and increasing the reliability of $\hat{\theta}$ (Camilli, 2018; Jessen et al., 2018; McNeish, 2023; McNeish & Wolf, 2020; Rhemtulla & Savalei, 2025), and participants with identical sum scores can have different $\hat{\theta}$ based on different patterns of item responses, thus providing theoretically more fine-grained distinctions between the respondents. Empirically, however, differences between CTT and IRT/FA scoring are often found to be minor (Lu et al., 2005; 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, unless the EIV correction is applied, as we will show.

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 (Bollen, 1989; Kline, 2023; Muthén, 2002). For example, consider the following LVM for the analysis of a treatment effect on test score data,

$$f(Y_{ij}) = \lambda_i(\theta_j + b_i) + \varepsilon_{ij} \tag{4}$$

$$\theta_j = \beta_0 + \beta_1 \text{treat}_j + \zeta_j \tag{5}$$

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon_i}) \tag{6}$$

$$\zeta_j \sim N(0, \sigma_\zeta) \tag{7}$$

in which the response Y to item i for person j is a function of latent person ability θ_j and item easiness parameters (item 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 an outcome score in a separate step.

Because θ_j is unobserved, constraints are necessary to identify the model and provide a scale for θ_j. Two standard approaches to model identification are the unit-loading approach, where λ_i is fixed to 1 for a single item (or all items, as in a Rasch model), or

[3] The ability of IRT to produce an interval scaling is a theoretical ideal that may or may not be met in any given empirical application, see e.g., Briggs and Domingue (2013); Briggs and Weeks (2009); Michell (1994, 1997); Reckase (2004); Schafer (2006) for various perspectives on this issue.

the unit-variance approach, where the total variance (or residual variance) of θ_j is fixed to 1. The item easiness parameters b_i can be identified by either excluding one item in a fixed effects approach (so that β_0 represents the performance of the average control respondent on this item), by fixing the mean of b_i to 0 in a random effects approach (De Boeck, 2008), or by fixing $\beta_0 = 0$.

These constraints resolve the scale indeterminacy of θ_j but are nonetheless arbitrary. For example, a single λ_i could be fixed to 2 instead of 1, yielding different point estimates and standard errors but providing identical fit to the observed data. Thus, when estimating causal effects on latent outcomes, indexing β_1 to the pooled standard deviation of θ_j (i.e., σ_ζ in the case without additional covariates in the model) is an attractive strategy to ground the interpretation of the model results. Accordingly, for the purposes of the present study, we target the following latent ATE in our analyses, following the notation of Stoetzer et al. (2024):

$$\frac{\mathbb{E}(\theta_j|T_j = 1) - \mathbb{E}(\theta_j|T_j = 0)}{\mathrm{SD}(\theta_j|T_j)}. \tag{8}$$

This approach has the important benefit of providing the same point estimate regardless of the chosen identification constraints, and is analogous to estimating a standardized effect size such as Cohen's *d* when the outcome variable is observed (Stoetzer et al., 2024).

$f()$ is a link function to allow for both linear and non-linear models. When all $\lambda_i = 1$, and the link function is logistic, i.e., $f(x) = \text{logit}(x)$, the LVM is equivalent to the One Parameter Logistic (1PL) Explanatory Item Response Model (EIRM; De Boeck, 2004, De Boeck & Wilson, 2016, Wilson et al., 2008). Note that the variance of the error term in the logistic model is fixed at $\frac{\pi^2}{3}$ for model identification (Breen et al., 2018; Mood, 2010). When λ_i are freely estimated and the link function is an identity, i.e., $f(x) = x$, 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 estimate associations among latent variables theoretically stripped of measurement error. LMVs therefore deattenuate estimates of standardized regression coefficients because, unlike σ_Y, σ_ζ is a consistent 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; Stoetzer et al., 2024; Zwinderman, 1991), suggesting that LVMs may provide more accurate tests of between-group differences such as causal treatment effects.

Model Assumptions

In addition to the standard causal inference assumptions of the stable unit treatment value assumption (SUTVA) and unconfoundedness of the treatment assignment, causal inference in latent variable contexts requires some additional assumptions. First, Equation 4 assumes full measurement invariance between the treatment and control groups. That is, other than the treatment effect on the latent trait, the items function equivalently between the groups. An example violation of this assumption could include "response shift," whereby treatment causes participants to interpret items differently such that differences in post-intervention scores reflect changes to item functioning rather than changes to the latent variable (Olivera-Aguilar & Rikoon, 2023). Stoetzer et al. (2024, p. 5) describe this assumption as "unconfounded measurement". More flexible models, such as multi-group estimation that allow for heteroskedasticity (Kim & Yoon, 2011) or multidimensional models that allow for response style effects (Deng et al., 2018) can relax these assumptions but are beyond the scope of the present study. Moreover, such approaches are relatively rare in applied causal inference (Soland, Edwards, & Talbert, 2024; Soland & Gilbert, 2025). We emphasize that while measurement invariance between treatment and control groups can and should be tested, such tests require item-level data, and therefore are difficult or impossible to assess in secondary analyses of sum or factor scores.

Figure 1 provides Directed Acyclic Graphs (DAGs) for the two-step and simultaneous estimation approaches and highlights a closely related assumption that is necessary for causal inference with latent variables. Namely, we must assume that the treatment effect on the individual item responses is fully mediated by θ_j, similar to the exclusion restriction assumption in instrumental variables estimation (Halpin & Gilbert, 2024; Stoetzer et al., 2024; VanderWeele & Vansteelandt, 2022). In other words, a treatment that improves θ_j is statistically equivalent to one that makes the items easier (Gilbert, Miratrix, et al., 2025; San Martín, 2016).[4]

Summary

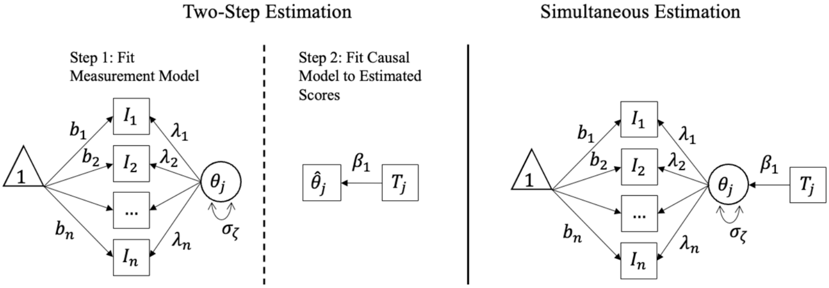
In sum, the applied researcher faces many choices in model selection when test score data are used as outcomes in a causal inference context: to use a one-step or two-step approach, to weight or not to weight the item responses in the construction of scores, to use CTT or IRT/FA, and so forth, as summarized in Table 1. 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 evaluation research by using Monte Carlo simulation and

[4] Under certain interpretations, the exclusion restriction assumption can be subsumed by the unconfounded measurement assumption because a direct effect on item performance beyond the latent trait implies that the items are functioning differentially between the treatment and control groups. However, alternative conceptions allow for item-specific treatment effects without invoking changes to the item parameters, but rather by introducing an additional item-specific treatment sensitivity term to the model (e.g., Gilbert, Himmelsbach, et al., 2025a, Figure F.1.)

an empirical application to examine the performance of several models under varying conditions of test score construction and model estimation.

Figure 1

Directed Acyclic Graphs for Two-Step and Simultaneous Estimation



Note. Squares indicate observed variables, hollow circles indicate latent variables. I_n are item responses, T_j is the treatment indicator, and β_1 represents the average treatment effect. σ_ζ is the residual standard deviation of the latent variable. See also Figure 1 in Stoetzer et al. (2024).

Table 1

Summary of Analysis Options for Estimating Treatment Effects on Latent Outcomes

Measurement Framework	Item Types	ATE Estimation Strategy	Weighting	Implementation
CTT	dichotomous, polytomous, continuous	Two-step	Equal	Generate sum score, run regression
IRT	dichotomous, polytomous	Two-step	Equal or Variable	Generate IRT score, run regression
FA	continuous	Two-step	Equal or Variable	Generate FA score, run regression
CTT	dichotomous, polytomous, continuous	Simultaneous	Equal	LMM on item responses[a]
IRT	dichotomous, polytomous	Simultaneous	Equal or Variable	EIRM on item responses
FA	continuous	Simultaneous	Equal or Variable	SEM on item responses

Note. CTT = Classical Test Theory, IRT = Item Response Theory, FA = Factor Analysis, LMM = Linear Mixed Model, EIRM = Explanatory Item Response Model, SEM = Structural Equation Model.

[a] The linear mixed model applicable to simultaneous CTT estimation is equivalent to the FA model with equal loadings when the item parameters are fixed; see Borsboom (2005) for a discussion of underlying equivalencies between models.

Method

Data Generating Process

The simulation and data analysis procedures are implemented in R. In total, we simulate 18,000 data sets (1,000 data sets per 18 data-generating conditions) and apply four analytic models—sum score, factor score, equal loading SEM, and variable loading SEM—to each, for a total of 72,000 results. We use 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 item characteristics, we fix 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 are drawn from $N(0 + \beta_1 \text{treat}_j, 1)$ and the item intercepts b_i are drawn from $N(0, 1)$. The latent variables are converted to continuous observed scores for each item using Equation 4. The residual *SD* for each item σ_{ε_i} is defined as $\sqrt{1 - \lambda_i^2}$ so that items with higher loadings have lower residual variances. The simulation factors include null, moderate, and large treatment effect sizes (0, 0.2, or 0.4 SDs on the latent trait, based on empirical distributions of effect sizes in education research, e.g., Kraft, 2020), 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 $x = 0, 0.15, 0.3$).

For each simulated data set, we estimate 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 item 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 ω 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.

We use `lm` to fit the OLS models, `lme4` to fit the FA models with equal loadings (Bates et al., 2015), and `lavaan` (Rosseel, 2012) to fit the FA models with variable loadings.[5] For the two-step approach using factor scores as outcomes, we use expected a posteriori (EAP) scores (Chapman, 2022; Lu et al., 2005; Muraki & Engelhard, 1985) as outcome variables. While there are many factor scoring options available in addition

[5] We use `lme4` to fit the equal-loading FA models because the syntax is more convenient than that of `lavaan`. We use MLE rather than the default REML in `lme4` so that the estimation procedures are identical across packages. Extensions to `lme4` such as `PLmixed` and `galamm` allow for more complex measurement models to be fit with `lme4` syntax (Rockwood & Jeon, 2019; Sørensen, 2024).

to EAP (Grice, 2001), such as maximum a posteriori (MAP), maximum likelihood, test characteristic curve (TCC, common in IRT, Baker et al., 2017), we use EAP scoring because it is among the most commonly used approaches and the Bayesian shrinkage of the EAP estimation reduces the *SD* of the observed scores, which is the primary cause of attenuation bias. For the EIV corrected two-step models, we calculate ω using the R package `psych` (Revelle & Condon, 2019). We collect the default *SE* provided by each model and we can estimate the true *SE* by calculating the *SD* of the point estimates.

Results

Figure 2 shows the mean 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 (as expected given that the standardized effect size is a ratio), and the bias is more severe when the average loadings are lower because lower average loadings translate 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.

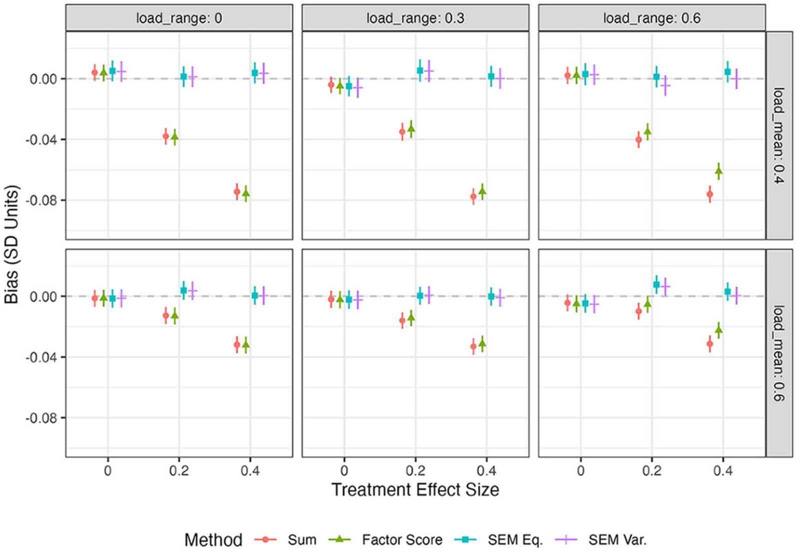
These results clearly illustrate that attenuation bias due to measurement error with standardized outcome variables is a more serious concern than the decision of whether to weight or not to weight the item responses. When we correct the two-step procedures for measurement error by dividing the coefficients by $\sqrt{\omega}$ as shown in Figure 3, we find that the performance of the sum score is indistinguishable from the LVMs.[6]

We include additional simulation results in Appendix A. In short, results show that differences between all models are minimal in terms of absolute precision (the *SD* of the point estimates), standard error calibration (the mean model-based *SE* as a proportion of

[6] Note that an EIV correction based on α will potentially overcorrect the factor score when the loadings are extremely variable. This occurs because the calculation of α assumes equal loadings and provides a lower bound on test reliability (McNeish & Wolf, 2020, p. 2292). When this assumption is not met, $\rho > \alpha$ so dividing by $\sqrt{\alpha}$ provides an overcorrection for the factor score model. Thus, if the range of loadings is large, a measure of reliability appropriate for models with variable factor loadings such as ω (Hayes & Coutts, 2020) should be applied, as we apply in our analyses. We also note that the concept of reliability is less straightforward in an IRT framework because the conditional standard error of measurement varies across the θ scale, e.g., Kim and Feldt (2010); Lockwood and McCaffrey (2014). However, marginal reliability in IRT provides a conceptually similar quantity that may be used for EIV corrections if desired, and ω functions well even with dichotomous items in moderate sample sizes (Padilla & Divers, 2016).

Figure 2

Estimated Bias by Method, Standardized Scores



Note. The points indicate the bias in standard deviation units, without the EIV correction applied to the two-step models. The bars indicate the Monte Carlo 95% CIs, calculated using the formula $\frac{s}{\sqrt{n}}$ where n is the number of simulation trials and s is the standard deviation of the point estimates.

the true *SE*), false positive rates, and statistical power, with the EIV correction applied to the two-step models. These results suggest 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. As a final sensitivity check, we rerun analogous simulations using dichotomous responses and IRT models rather than continuous responses and FA models. We find essentially identical results to those reported here, suggesting that our findings are consistent across multiple specifications and outcome item types. We include the full IRT simulation analysis results in our supplement.

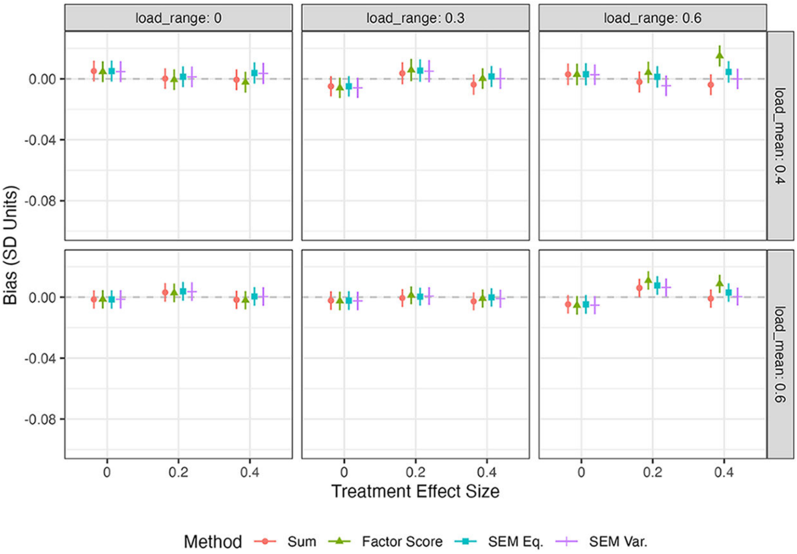
Empirical Application

Data Source

To illustrate how the issues of scoring and model selection can play out in practice, we use a selection of empirical datasets from RCTs containing item-level outcome data from Gilbert, Himmelsbach, et al. (2025a). The authors examine 75 datasets from 48 RCTs

Figure 3

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



Note. The points indicate the bias in standard deviation units, with the EIV correction applied to the two-step models. The bars indicate the Monte Carlo 95% CIs, calculated using the formula $\frac{s}{\sqrt{n}}$ where n is the number of simulation trials and s is the standard deviation of the point estimates.

with item-level outcome data to examine models for item-level heterogeneous treatment effects, in which treatments may uniquely impact each item of an outcome measure. Here, we take a random sample of 10 datasets to illustrate the results of different analytic approaches to estimating average treatment effects across a range of contexts. Table 2 provides a summary of each dataset. The datasets cover a range of geographic regions, outcome measures, and show a wide range of estimated reliability with ω ranging from 0.43 to 0.95. In contrast to the simulation, we cannot know the true value of the treatment effect on the latent trait in the empirical data. However, 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 choices.

Analytic Models

We apply eight estimators to produce treatment effects from each dataset. Because all item responses are dichotomous, we use logistic IRT models instead of the linear SEM. For two-step approaches, we calculate the mean score, 1PL IRT score, and 2PL IRT score, and regress the resulting scores on the treatment indicator and standardize the results to

Table 2

Studies Included in Our Empirical Analysis

ID	Citation	Location	Outcome	*N*	*I*	ω
1	Bruhn et al. (2016)	Brazil	Financial Literacy	16852	10	0.57
2	Kim et al. (2024)	USA	Reading Comprehension	1335	29	0.86
3	Kim et al. (2021)	USA	Vocabulary	2588	24	0.79
4	Romero et al. (2020)	Liberia	Raven’s Progressive Matrices	3510	10	0.63
5	Duflo et al. (2024)	Ghana	English	27201	21	0.89
6	Carpena (2024)	India	Health Knowledge	839	21	0.75
7	Luo et al. (2019)	China	Parenting Beliefs	449	11	0.43
8	Lyall et al. (2020)	Afghanistan	Pro-government attitudes	1853	9	0.49
9	Banerjee et al. (2017)	India	Math	9193	30	0.93
10	Banerjee et al. (2017)	India	Math	5356	30	0.95

Note. *N* = number of subjects, *I* = number of items, ω = estimated reliability of the outcome measure.

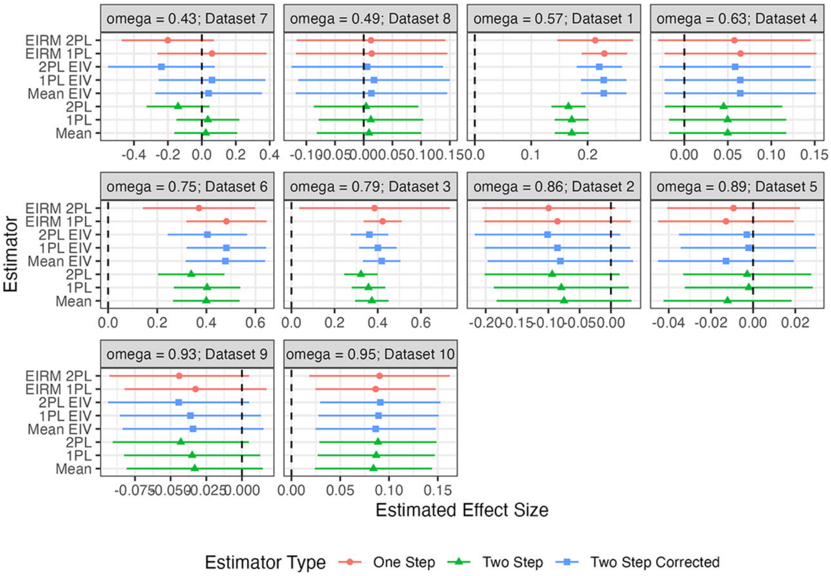
calculate Cohen’s *d*. We then apply the EIV correction to these three estimates, dividing the estimated ATE by $\sqrt{\omega}$. (Our supplemental simulations of IRT models show that the classical EIV correction works well even though single-number estimates of reliability are less common in IRT frameworks where precision varies over the range of the latent trait, see Nicewander, 2018; Raju et al., 2007). We use mean scores instead of sum scores because there is some missing item response data. For one-step approaches, we estimate 1PL and 2PL EIRMs that allow for a treatment effect directly on the latent trait.

Results

Figure 4 shows the point estimates and 95% CIs for the standardized treatment effect from the eight estimators applied to our 10 datasets. The datasets are ordered by ω from lowest to highest. When ω is high (datasets 2, 5, 9, 10), differences between the estimators are minimal, as expected. Datasets with moderate ω (datasets 1, 4, 6, 3) most clearly mirror the simulation results, showing estimates from two-step models generally lower than alternative approaches, and minimal differences between EIV corrected estimates and estimates from the one-step models. In dataset 8, the ATE is near 0 and all estimators are very close to the null value. Only when ω is very low, as in dataset 7 ($\omega = .43$), do we see meaningful differences between weighted and unweighted estimators, with the 2PL approaches yielding negative point estimates and the 1PL approaches yielding positive point estimates. Taken as a whole, these results again suggest that once EIV corrections are applied, differences between estimators are likely to be minor in all but the most extreme cases.

Figure 4

Estimated Treatment Effects for 10 RCT Datasets



Note. The points indicate the estimated treatment effect size in standard deviation units. The bars indicate the model-based 95% CIs. 1PL and 2PL indicate IRT-based scores are derived from one-parameter logistic and two-parameter logistic models, respectively. The EIRM is equivalent to an SEM with a logistic link function.

Discussion

Because psychometric outcome measures 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 outcome data with different properties, the bias is substantial when treatment effect sizes are high, as predicted by Classical Test Theory. However, when the EIV correction is applied and the standardized coefficients are divided by $\sqrt{\omega}$, 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. When left unaddressed, such bias could lead researchers and policymakers to erroneous conclusions about the efficacy of interventions.

As a concrete example of how attenuation bias could affect substantive results, consider meta-analyses that pool the effects of interventions on test score outcomes across studies. Even if the true underlying effect is equal across all studies, when the outcome measures are of varying reliability, the estimated effect sizes will differ due to attenuation bias, even as the participant and study sample sizes grow to infinity (Borenstein et al., 2009). Thus, failing to adjust standardized effect sizes for measurement error may lead to spurious conclusions about treatment heterogeneity. The degree to which such issues may be related to the ongoing "replication crisis" in psychology and other fields is an open question (Open Science Collaboration, 2015), but it seems plausible that variation in measurement practices may play a role in explaining variation in conclusions across studies (Flake et al., 2017; Flake & Fried, 2020; Pedersen et al., 2025).

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 (McNeish, 2022; McNeish & Wolf, 2020), rather than the first-order issue of attenuation of standardized coefficients for measurement error in the outcome variable (Shear & Briggs, 2024; Widaman & Revelle, 2023). 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 in applied program evaluation.

These results should not detract from other uses of LVMs. 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, 2024a), 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 differentially 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., 2024; Gilbert, 2024b; Gilbert, Himmelsbach, et al., 2025a; Gilbert et al., 2023; Sales et al., 2021), differential growth by item type (Briggs, 2021; Gilbert et al., 2024; Naumann et al., 2014), or the appropriate interpretation of interaction effects (Domingue, Kanopka, Trejo, et al., 2024; Gilbert, Miratrix, et al., 2025). However, when all respondents answer 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, despite arguments that the sum score can be a suboptimal choice (in general) 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 evidence for the importance of EIV corrections in two-step analyses of standardized test score outcome variables, several limitations merit consideration. For example, 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, Himmelsbach, et al., 2025a; Gilbert et al., 2024; Gilbert, Miratrix, et al., 2025; Kuhfeld & Soland, 2022, 2023; Mayer, 2019; Miratrix et al., 2021; Rabbitt, 2018; Soland, 2022, 2023; Soland et al., 2023).

A related issue is measurement error in covariates, which we did not explore in this study. In theory, in an RCT, any bias induced by covariate measurement error will affect treatment and control groups equally and thus should not affect estimation of the ATE (Lockwood & McCaffrey, 2014). In observational studies, however, covariate measurement error can lead to biases when the covariates do not fully control for relevant differences between groups (Cook et al., 2009; Sengewald & Pohl, 2019; Sengewald et al., 2019). Factor models with latent covariates and outcomes are easily estimable in `lavaan` when the indicators are continuous, however, software options for categorical responses common in the social sciences are currently less widely used in R, though recent developments such as `galamm` (Sørensen, 2024) and `EffectLiteR` (Mayer et al., 2016, 2020; Sengewald & Mayer, 2024) may provide attractive options. We view exploration of how measurement error in *both* covariates and outcomes influences results in experimental and observational contexts to be a promising avenue for future research.

In conclusion, results of causal analyses of psychometric outcome 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 or use 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 evaluation research.

Funding: The author has no funding to report.

Acknowledgments: The author wishes to thank Andrew Ho, Ben Domingue, Derek Briggs, and two anonymous reviewers for their helpful comments on drafts of this paper.

Competing Interests: The author has declared that no competing interests exist.

Data Availability: A replication toolkit is available at Gilbert, 2025. The empirical datasets and study materials from Gilbert, Himmelsbach, et al. (2025b) are available at https://doi.org/10.7910/DVN/C4TJCA, and also in the Item Response Warehouse (IRW, Domingue et al., 2024), with the prefix `gilbert_meta`: https://redivis.com/datasets/as2e-cv7jb41fd.

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Data. (Gilbert, Himmelsbach, et al., 2025b)
- Study materials. (Gilbert, Himmelsbach, et al., 2025b)
- Replication toolkit. (Gilbert, 2025)

References

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Benjamin, W. (2024). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials, 1–22. *Journal of Research on Educational Effectiveness.* https://doi.org/10.1080/19345747.2024.2361337

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Asher, H. B. (1974). Some consequences of measurement error in survey data. *American Journal of Political Science, 18*(2), 469–485.

Baker, F. B., & Kim, S.-H. (2017). The test characteristic curve. *The basics of Item Response Theory using R* (pp. 55–67). Springer.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives, 31*(4), 73–102. https://doi.org/10.1257/jep.31.4.73

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

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197–211. https://doi.org/10.1111/J.1745-3984.1997.Tb00515.X

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* John Wiley & Sons.

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

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., & van Smeden, M. (2018). Measurement error is often neglected in medical literature: A systematic review. *Journal of Clinical Epidemiology, 98*, 89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and understanding logits, probits, and other nonlinear probability models. *Annual Review of Sociology, 44*, 39–54. https://doi.org/10.1146/annurev-soc-073117-041429

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89–118. https://doi.org/10.1080/08957340801926086

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. https://doi.org/10.3102/1076998613508317

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://ideas.repec.org/a/tpr/edfpol/v4y2009i4p384-414.html

Bruhn, M., de Souza Leão, L., Legovini, A., Marchetti, R., & Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics, 8*(4), 256–295. https://doi.org/10.1257/app.20150149

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

Carpena, F. (2024). Entertainment-education for better health: Insights from a field experiment in India. *Journal of Development Studies, 60*(5), 745–762. https://doi.org/10.1080/00220388.2024.2312832

Carroll, R. J., Delaigle, A., & Hall, P. (2009). Nonparametric prediction in measurement error models. *Journal of the American Statistical Association, 104*(487), 993–1003. https://doi.org/10.1198/jasa.2009.tm07543

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

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

Chapman, R. (2022). Expected a posteriori scoring in PROMIS®. *Journal of Patient-Reported Outcomes, 6*, Article 59. https://doi.org/10.1186/s41687-022-00464-9

Christensen, K. B. (2006). From Rasch scores to regression. *Journal of Applied Measurement, 7*(2), 184–191.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315. https://doi.org/10.1037/a0033805

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research, 44*(6), 828–847. https://doi.org/10.1080/00273170903333673

Cox, K., & Kelcey, B. (2019). Optimal design of cluster-and multisite-randomized studies using fallible outcome measures. *Evaluation Review, 43*(3–4), 189–225. https://doi.org/10.1177/0193841X19870878

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

De Boeck, P. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer Science & Business Media.

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

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

Deng, S., E. McCarthy, D., E. Piper, M., B. Baker T., & Bolt, D. M. (2018). Extreme response style and the measurement of intra-individual variability in affect. *Multivariate Behavioral Research, 53*(2), 199–218. https://doi.org/10.1080/00273171.2017.1413636

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11 Suppl. 3), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Domingue, B., Braginsky, M., Caffrey-Maffei, L. A., Gilbert, J., Kanopka, K., Kapoor, R., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2024). *Solving the problem of data in psychometrics: An introduction to the Item Response Warehouse (IRW).* PsyArXiv. 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., Trejo, S., Rhemtulla, M., & Tucker-Drob, E. M. (2024). 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, 29*(6), 1164–1179. https://doi.org/10.1037/met0000532

Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental evidence on four policies to increase learning at scale. *Economic Journal, 134*(661), 1985–2008. https://doi.org/10.1093/ej/ueae003

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.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fuller, W. A., & Hidiroglou, M. A. (1978). Regression estimation after correcting for attenuation. *Journal of the American Statistical Association, 73*(361), 99–104. https://doi.org/10.1080/01621459.1978.10480011

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time* (pp. 1–17). Department of Statistics, Columbia University.

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

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

Gilbert, J. B. (2025). *ResearchBox #2289 – 'How measurement affects causal inference'* [Replication toolkit]. ResearchBox. https://researchbox.org/2289.

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025a). Estimating heterogeneous treatment effects with item-level outcome data: Insights from Item Response Theory. *Journal of Policy Analysis and Management,* 1–34. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025b). *"Replication data for: Estimating heterogeneous treatment effects with item-level outcome data: Insights from Item Response Theory"* [Data, Study Materials]. *Harvard Dataverse.* https://doi.org/10.7910/DVN/C4TJCA.

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

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

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

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *REVSTAT-Statistical Journal, 8*(1), 57–80. https://doi.org/10.57805/revstat.v8i1.90

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450. https://doi.org/10.1037/1082-989X.6.4.430

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

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*(4), 373–378. https://doi.org/10.1177/014662168200600401

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105–131. https://doi.org/10.3102/00346543071001105

Hayes, A. F., & Coutts, J. J. (2020). Use Omega rather than Cronbach's Alpha for estimating reliability. but… *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.2307/1164588

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. https://doi.org/10.1177/0146621615585851

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ, 340*, Article c2289. https://doi.org/10.1136/bmj.c2289

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

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. https://doi.org/10.1177/0146621616664046

Jackson, P. H. (1973). The estimation of true score variance and error variance in the classical test theory model. *Psychometrika, 38*(2), 183–201. https://doi.org/10.1007/BF02291113

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. https://doi.org/10.1177/0194599818760528

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology, 113*(1), 3–26. https://doi.org/10.1037/edu0000465

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review, 11*(2), 179–188. https://doi.org/10.1007/s12564-009-9062-8

Kim, J. S., Gilbert, J. B., Relyea, J. E., Rich, P., Scherer, E., Burkhauser, M. A., & Tvedt, J. N. (2024). Time to transfer: Long-term effects of a sustained and spiraled content literacy intervention in the elementary grades. *Developmental Psychology, 60*(7), 1279–1297. https://doi.org/10.1037/dev0001710

Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis, 27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

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

Kuhfeld, M., & Soland, J. (2023). Scoring assessments in multisite randomized control trials: Examining the sensitivity of treatment effect estimates to measurement choices. *Psychological Methods.* https://doi.org/10.1037/met0000633

Lee, W.-C., & Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth, & D. J. Hughes (Eds.),*Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 639–673). Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch21

Lewis, C. (2006). 2 selected topics in classical test theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 29–43). https://doi.org/10.1016/S0169-7161(06)26002-4

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology, 127*(4), 864–874. https://doi.org/10.1093/oxfordjournals.aje.a114870

Liu, Y., & Pek, J. (2024). Summed versus estimated factor scores: Considering uncertainties when using observed scores. *Psychological Methods.* https://doi.org/10.1037/met0000644

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/10.3102/107699861350940

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

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

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on irt scores. *Structural Equation Modeling, 12*(2), 263–277. https://doi.org/10.1207/s15328007sem1202_5

Luo, R., Emmers, D., Warrinnier, N., Rozelle, S., & Sylvia, S. (2019). Using community health workers to deliver a scalable integrated parenting program in rural China: A cluster-randomized controlled trial. *Social Science & Medicine, 239*, Article 112545. https://doi.org/10.1016/j.socscimed.2019.112545

Lyall, J., Zhou, Y.-Y., & Imai, K. (2020). Can economic assistance shape combatant support in wartime? Experimental evidence from Afghanistan. *American Political Science Review, 114*(1), 126–143. https://doi.org/10.2139/ssrn.3026531

Mayer, A. (2019). Causal effects based on latent variable models. *Methodology, 15*(S1), 15–28. https://doi.org/10.1027/1614-2241/a000174

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR Approach for Analyzing Average and Conditional Effects. *Multivariate Behavioral Research, 51*(2–3), 374–391. https://doi.org/10.1080/00273171.2016.1151334

Mayer, A., Zimmermann, J., Hoyer, J., Salzer, S., Wiltink, J., Leibing, E., & Leichsenring, F. (2020). Interindividual differences in treatment effects based on structural equation models with latent variables: An EffectLiteR tutorial. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(5), 798–816. https://doi.org/10.1080/10705511.2019.1671196

McNeish, D. (2022). Limitations of the sum-and-alpha approach to measurement in behavioral research. *Policy Insights from the Behavioral and Brain Sciences, 9*(2), 196–203. https://doi.org/10.1177/23727322221117144

McNeish, D. (2023). 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. https://doi.org/10.3758/s13428-020-01398-0

Michell, J. (1994). Numbers as quantitative relations and the traditional theory of measurement. *British Journal for the Philosophy of Science, 45*(2), 389–406. 10.1093/bjps/45.2.389

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

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. https://doi.org/10.1080/19345747.2020.1831115

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Muraki, E., & Engelhard, G., Jr. (1985). Full-information item factor analysis: Applications of eap scores. *Applied Psychological Measurement, 9*(4), 417–430. https://doi.org/10.1177/014662168500900411

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research.* Oxford University Press.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81–117. https://doi.org/10.2333/bhmk.29.81

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. https://doi.org/10.1111/jedm.12051

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351–362. https://doi.org/10.1037/met0000132

Olivera-Aguilar, M., & Rikoon, S. H. (2023). Intervention effect or measurement artifact? Using invariance models to reveal response-shift bias in experimental studies. *Journal of Research on Educational Effectiveness,* 1–29. https://doi.org/10.1080/19345747.2023.2284768

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

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

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: Coefficient Omega confidence intervals in the current literature. *Educational and Psychological Measurement, 76*(3), 436–453. https://doi.org/10.1177/0013164415593776

Pedersen, A. P., Kellen, D., Mayo-Wilson, C., Davis-Stober, C. P., Dunn, J. C., Khan, M. A., Stinchcombe, M. B., Kalish, M. L., Tentori, K., & Haaf, J. (2025). Discourse on measurement. *Proceedings of the National Academy of Sciences of the United States of America, 122*(5), Article e2401229121. https://doi.org/10.1073/pnas.2401229121

Rabbitt, M. P. (2018). Causal inference with latent variables from the Rasch model as outcomes. *Measurement, 120*, 193–205. https://doi.org/10.1016/j.measurement.2018.01.044

Raju, N. S., Price, L. R., Oshima, T., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*(3), 169–180. https://doi.org/10.1177/0146621606291569

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1), 117–120.

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

Rhemtulla, M., & Savalei, V. (2025). Estimated factor scores are not true factor scores. *Multivariate Behavioral Research, 60*(3), 598–619. https://doi.org/10.1080/00273171.2024.2444943

Rockwood, N. J., & Jeon, M. (2019). Estimating complex measurement and growth models using the R package PLmixed. *Multivariate Behavioral Research, 54*(2), 288–306. https://doi.org/10.1080/00273171.2018.1516541

Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review, 110*(2), 364–400. https://doi.org/10.1257/aer.20181478

Rosenbaum, P. (2017). *Observation and experiment: An introduction to causal inference.* Harvard University Press.

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). *The effect of an intelligent tutor on performance on specific posttest problems* (pp. 206–215). Proceedings of the 14th International Conference on Educational Data Mining (EDM21), Paris, France, June 29–July 2, 2021. International Educational Data Mining Society.

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 2, pp. 127–150). Chapman; Hall/CRC.

Schafer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, Research, and Evaluation, 11*(1), Article 4. https://doi.org/10.7275/xjkz-7n67

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution, 1*(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

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*, Article 24. https://doi.org/10.1186/1471-2288-10-24

Sengewald, M.-A., & Mayer, A. (2024). Causal effect analysis in nonrandomized data with latent variables and categorical indicators: The implementation and benefits of EffectLiteR. *Psychological Methods, 29*(2), 287–307. https://doi.org/10.1037/met0000489

Sengewald, M.-A., & Pohl, S. (2019). Compensation and amplification of attenuation bias in causal effect estimates. *Psychometrika, 84*(2), 589–610. https://doi.org/10.1007/s11336-019-09665-6

Sengewald, M.-A., Steiner, P. M., & Pohl, S. (2019). When does measurement error in covariates impact causal effect estimates? Analytic derivations of different scenarios and an empirical illustration. *British Journal of Mathematical and Statistical Psychology, 72*(2), 244–270. https://doi.org/10.1111/bmsp.12146

Shear, B. R., & Briggs, D. C. (2024). Measurement issues in causal inference. *Asia Pacific Education Review, 25*(3), 719–731. https://doi.org/10.1007/s12564-024-09942-9

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

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. https://doi.org/10.1177/0956797611417632

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. https://doi.org/10.1177/00131644211007551

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, 17*(2), 391–421. https://doi.org/10.1080/19345747.2023.2195413

Soland, J., Edwards, K., & Talbert, E. (2024). When should evaluators lose sleep over measurement? Toward establishing best practices. *Journal of Research on Educational Effectiveness,* 1–33. https://doi.org/10.1080/19345747.2024.2344011

Soland, J., & Gilbert, J. B. (2025). *Does socially desirable responding increase after an intervention? Implications for estimating treatment effects*. PsyArXiv. https://doi.org/10.31234/osf.io/ujx4n_v1

Soland, J., Johnson, A., & Talbert, E. (2023). Regression discontinuity designs in a latent variable framework. *Psychological Methods, 28*(3), 691–704. https://doi.org/10.1037/met0000453

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. https://doi.org/10.1037/met0000506

Sørensen, Ø. (2024). Multilevel semiparametric latent variable modeling in R with "galamm". *Multivariate Behavioral Research, 59*(5), 1098–1105. https://doi.org/10.1080/00273171.2024.2385336

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*(6), 605–627. https://doi.org/10.3102/1076998616655442

Stoetzer, L., Zhou, X., & Steenbergen, M. (2024). Causal inference with latent outcomes. *American Journal of Political Science, 62*(2), 624–640. https://doi.org/10.1111/ajps.12871

Thissen, D., & Wainer, H.. (Eds.). (2001). An overview of test scoring. *Test scoring* (pp. 13–32). Taylor & Francis. https://doi.org/10.4324/9781410604729

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*, Article 102486. https://doi.org/10.1016/j.ssresearch.2020.102486

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(8), 8–14.

VanderWeele, T. J., & Vansteelandt, S. (2022). A statistical test to reject the structural interpretation of a latent factor model. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 84*(5), 2032–2054. https://doi.org/10.1111/rssb.12555

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,* Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

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. https://doi.org/10.3758/s13428-022-01849-w

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Hogrefe & Huber Publishers.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement, 72*(3), 453–468. https://doi.org/10.1177/0013164411419846

Ye, F. (2016). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology, 69*(1), 43–61. https://doi.org/10.1111/bmsp.12058

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56*(4), 589–600. https://doi.org/10.1007/BF02294492

Appendix

Additional Simulation Results

All figures include EIV corrections for the two-step scores. The figure for statistical power only includes true effect sizes of 0.2 because near-ceiling levels of power are achieved at an effect size of 0.4.

Figure A.1.

Estimated Standard Errors by Method

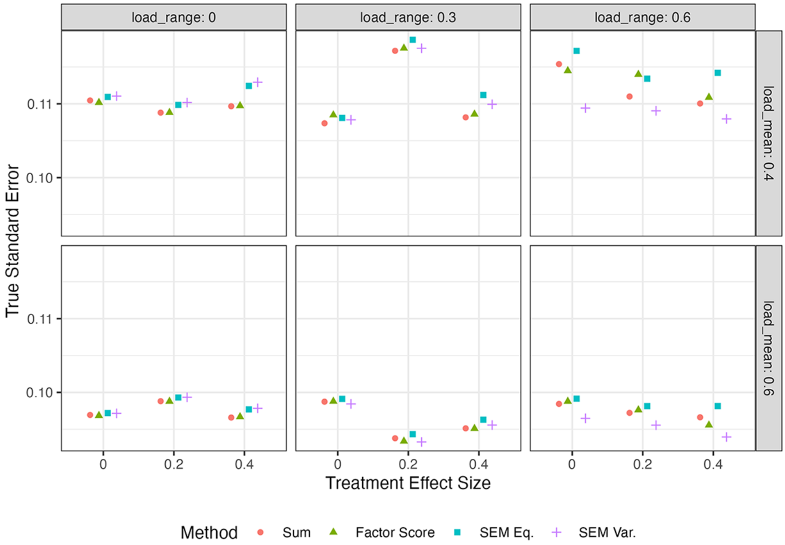


Figure A.2.

Estimated Standard Error Calibration by Method

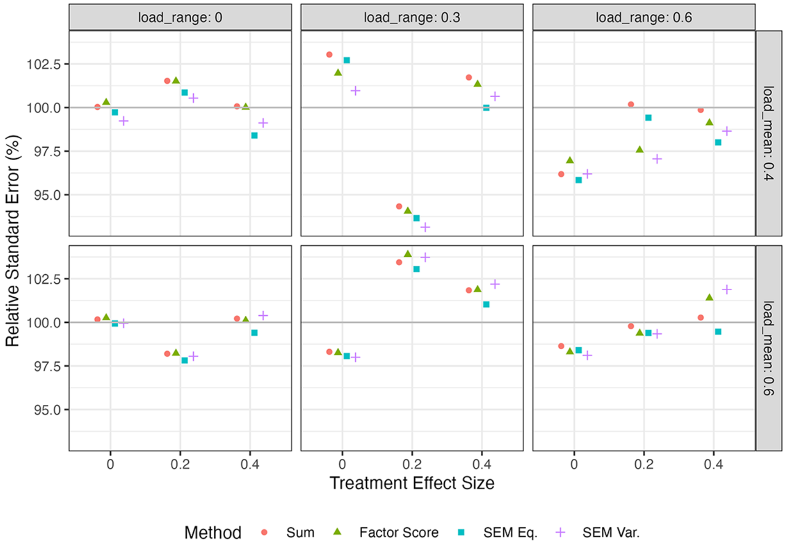


Figure A.3.

Estimated False Positive Rates by Method

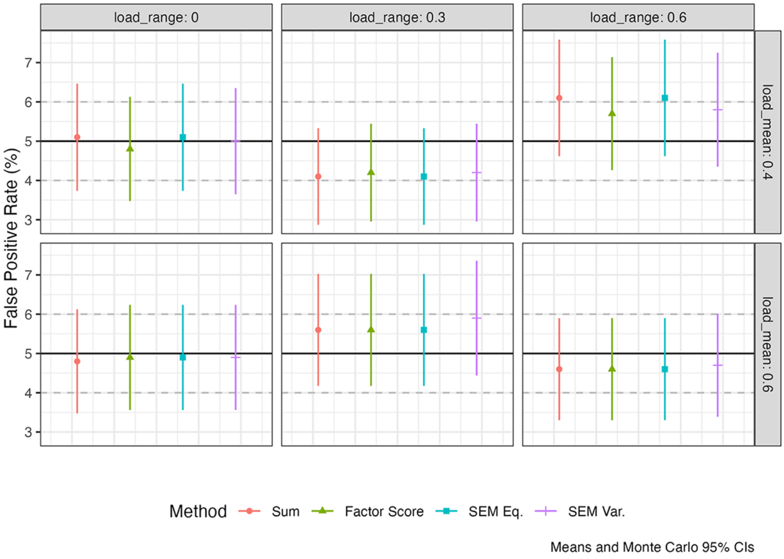
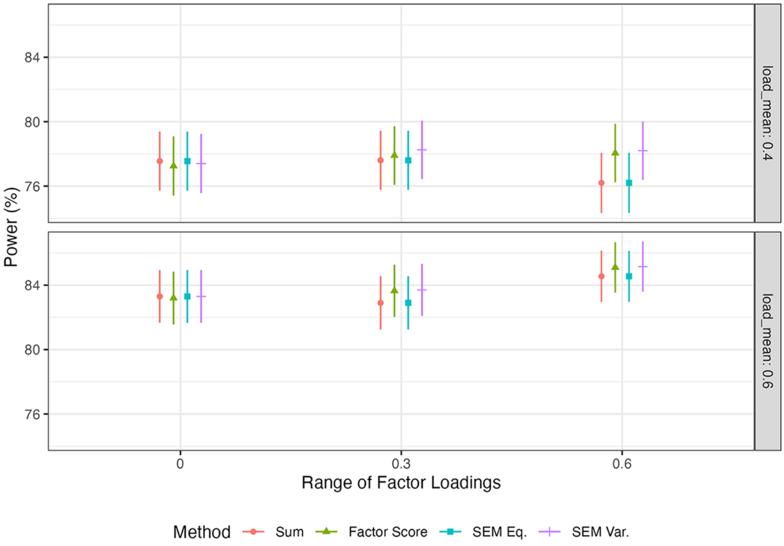


Figure A.4.

Estimated Power by Method


load_mean: 0.4
load_mean: 0.6
Power (%)
Range of Factor Loadings
Method
Sum
Factor Score
SEM Eq.
SEM Var.