Causal Inference for Experiments with Latent Outcomes: Key Results and Their Implications for Design and Analysis

Jiawei Fu* Donald P. Green†

April 20, 2025

Abstract

How should we analyze randomized experiments in which the main outcome is measured in multiple ways and each measure contains some degree of error? Since Costner (1971) and Bagozzi (1977), methodological discussions of experiments with latent outcomes have reviewed the modeling assumptions that are invoked when the quantity of interest is the average treatment effect (ATE) of a randomized intervention on a latent outcome that is measured with error. Many authors have proposed methods to estimate the ATE when multiple measures of an outcome are available. Despite this extensive literature, social scientists rarely use these modeling approaches when analyzing experimental data, perhaps because the surge of interest in experiments coincides with increased skepticism about the modeling assumptions that these methods invoke. The present paper takes a fresh look at the use of latent variable models to analyze experiments. Like the skeptics, we seek to minimize reliance on ad hoc assumptions that are not rooted in the experimental design and measurement strategy. At the same time, we think that some of the misgivings that are frequently expressed about latent variable models can be addressed by modifying the research design in ways that make the underlying assumptions defensible or testable. We describe modeling approaches that enable researchers to identify and estimate key parameters of interest, suggest ways that experimental designs can be augmented so as to make the modeling requirements more credible, and discuss empirical tests of key modeling assumptions. Simulations and an empirical application illustrate the gains in terms of precision and robustness.

*Postdoctoral Associate, Yale University (jf3739@nyu.edu)

†Burgess Professor of Political Science, Columbia University (dpg2110@columbia.edu).

We thank David Broockman, Alex Coppock, Macartan Humphreys, Josh Kalla, Patrick Liu, Ryan Moore and participants of the UC San Diego Political Methodology and American Politics Speaker Series for their helpful comments on earlier drafts. We welcome feedback on this draft. A preliminary R package can be found at Github.

1 Introduction

Social scientists often seek to estimate the average causal effect of interventions and increasingly rely on experimental designs to do so convincingly. The statistical literature on experimental design and analysis has grown markedly in recent years, with noteworthy advances in asymptotic theory (Li and Ding, 2017), modeling (Sävje et al., 2021), and estimation (Su et al., 2023). However, one topic of special concern to social scientists has largely escaped attention, even from otherwise comprehensive textbooks: imperfect measurement of experimental outcomes. Angrist and Pischke (2009), Gerber and Green (2012), and Imbens and Rubin (2015) offer no sustained formal treatment of outcome measurement or how to analyze experiments in which outcomes are measured in more than one way. The models that these textbooks present implicitly assume that the observed outcome is a manifestation of the true underlying potential outcomes of interest.

In the social sciences, however, there is often slippage between the outcomes of interest and the proxy outcome measures at hand. Constructs such as economic inequality, press freedom, political violence, corruption, and post-materialism are just a few examples of latent variables that are thought to be measured imperfectly, despite the sustained efforts of researchers.

Our paper builds on recent attempts to formalize the estimation challenges that may arise when latent outcomes are measured with error. Like Stoetzer et al. (2022), we use potential outcomes notation to unify the treatment of experimental design and outcome measurement. We make four contributions, each of which has important implications for research design and estimation. First, building on a largely forgotten literature dating at least to Costner (1971) and Alwin and Tessler (1974), we show how an intervention's average treatment effect on a latent outcome can be identified and estimated without "standardizing" the latent outcome in ways that make the units of measurement sample-specific. Adapting ideas from prior work by Bagozzi (1977), Sorbom (1981), and Kano (2001) to a potential outcomes framework, we propose efficient and easily implemented estimators for the experimental intervention's ATE on the latent outcome. Second, we show analytically and through simulation how multiple measures of a latent outcome can improve the precision with which the ATE is estimated. These results highlight an underappreciated practical trade-off between gathering more observations and gathering additional outcome measures. Third, we show how certain data collection strategies can help satisfy what would otherwise be unrealistic measurement assumptions, such as the stipulation that an observed measure represents a linear function of the underlying construct of interest plus measurement error.[1] We are by no means the first to note that "invalid" measurements threaten unbiased inference or that redundant outcome measurements can improve precision; our contribution is to bring a coherent analytic framework to the discussion of experimental research design and outcome measurement so as to clarify the value of research designs that allocate resources to the collection of multiple measurements of a latent outcome. Fourth, we show how latent outcome models may be viewed as nested alternatives to seemingly unrelated regression (SUR), allowing researchers to assess empirically whether the constraints imposed by the latent outcome model are consistent with the data. If not, researchers can fall back on the more agnostic SUR approach in which each individual outcome measure is

[1] Our approach to data collection also reduces reliance on nonlinear models that assume constant treatment effects among all subjects, such as the hierarchical IRT model proposed by Stoetzer et al. (2022). We offer design recommendations that allow the analyst to stay within a simpler linear modeling framework that accommodates heterogeneous treatment effects but recognize that researchers sometimes have no control over the available outcome measures and may need to take a nonlinear modeling approach.

regressed on the treatment.

This paper is structured as follows. We begin by presenting a potential outcomes model that defines the target of inference and allows for some degree of slippage between the latent outcome we seek to measure and the observed measures at our disposal. Next, we show formally the conditions under which the average treatment effect on the posited latent outcome is identified. When latent outcomes are measured with random error, this average treatment effect may be estimated consistently, but precision is enhanced when the researcher gathers multiple outcome measures. When measurement errors are systematic – especially when mismeasurement operates differently for subjects assigned to treatment or control – methods that are typically used to analyze experiments may no longer yield unbiased estimates. We show how identification problems may be diagnosed empirically and addressed by diversifying the portfolio of measures and adopting an estimator that leverages the additional measures in a theoretically informed way. A simulated example helps explicate the mechanics of this approach. In order to show its relevance to applied empirical work, we reanalyze the Kalla and Broockman (2020) experiment, which features a diverse array of outcome measures that improve precision and facilitate robustness checks.

2 Framework

Consider a sample of n units drawn from a population of size N. Let Z_i denote the treatment assignment for unit i. Generalizing the potential outcomes framework, we assume two potential latent variables: η_i^1 is the latent outcome that would be realized for subject i if the treatment were administered and η_i^0 is the latent outcome that would be realized for subject i in the absence of treatment. These latent variables are not directly observed; think of them instead as abstract concepts, such as authoritarianism or corruption, that may be measured with some degree of error. Suppose that there are J observed measures of the latent outcome for each subject. Let the j^{th} outcome measure for unit i be $Y_{ij} = \lambda_j \eta_i + \epsilon_{ij} = \lambda_j[Z_i\eta_i^1 + (1 - Z_i)\eta_i^0] + \epsilon_{ij}$, where $\mathbb{E}[\epsilon_i] = 0$.[2] It is also sometimes useful for us to define $\epsilon'_{ij} = \frac{1}{\lambda_j}\epsilon_{ij}$, and thus $Y_{ij} = \lambda_j[Z_i\eta_i^1 + (1 - Z_i)\eta_i^0 + \epsilon'_{ij}]$. Figure 1 illustrates the relationship among the observed and unobserved variables. The observed variables are depicted inside squares, and unobserved variables are depicted inside circles. The latent outcome η_i can be represented as $\eta_i = \mathbb{E}\eta_i^0 + \tau Z_i + \zeta_i$, where ζ_i is the idiosyncratic disturbance: $\zeta_i = \eta_i^0 - \mathbb{E}\eta_i^0 + Z_i[(\eta_i^1 - \mathbb{E}\eta_i^1) - (\eta_i^0 - \mathbb{E}\eta_i^0)]$. One can verify that this term has mean zero by construction: $\mathbb{E}\zeta_i = 0$.

We are interested in the causal effect of treatment on the latent variable, which Stoetzer et al. (2022) call the latent treatment effect (LTE). Define the individual-level LTE as $\tau_i = \eta_i^1 - \eta_i^0$. Because of the fundamental problem of causal inference, we focus on the average effect. The average LTE for a given set of subjects is $\tau = \frac{1}{n}\sum_{i=1}^n \eta_i^1 - \frac{1}{n}\sum_{i=1}^n \eta_i^0 = \frac{1}{n}\sum_{i=1}^n \tau_i$; the superpopulation average treatment effect is $\mathbb{E}[\eta_i^1] - \mathbb{E}[\eta_i^0]$.

Our identification strategy is rooted in a set of basic assumptions about the experimental design and the manner in which the latent variables are measured:

Assumption 1 (Causal Framework for a Latent Outcome Variable)**.** *In the Causal Framework for a Latent Outcome Variable, we assume* $\forall j = 1, 2, ..., J$*, and* $i = 1, 2, ..., n$*,*

A. Valid Measurement: $\lambda_j \neq 0$.

[2]This assumption can be relaxed to any constant since outcomes can be mean-centered.

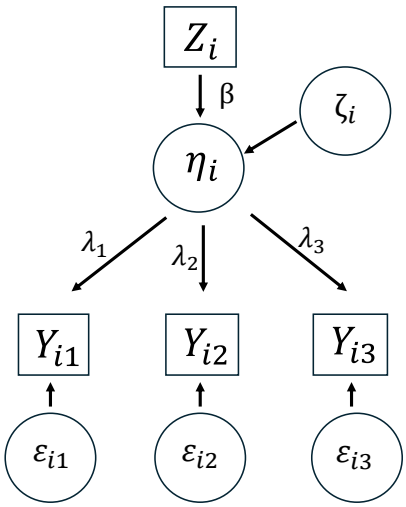

Figure 1: Graphical Depiction of an Experimental Design in which a Latent Outcome is Measured Linearly by Three Outcomes, Each Measured with Error.

B. Z_i *is randomly assigned:* $Var(Z_i) \neq 0$ *and* $\{\eta_i^0, \eta_i^1\} \perp Z_i$
C. Z_i *is excludable:* $\epsilon_{ij} \perp Z_i$
D. Measurement error is independent of potential outcomes: $\{\eta_i^0, \eta_i^1\} \perp \epsilon_{ij}$
E. Stable unit treatment value assumption (SUTVA) (Rubin, 1980):
a. If $Z_i = Z_i'$*, then* $\eta_i(Z_1, Z_2, ..., Z_i, ..., Z_n) = \eta_i(Z_1, Z_2, ..., Z_i', ..., Z_n)$
b. If $Z_i = z$*, then* $\eta_i = \eta_i(z)$, $\forall i$ *and* $z \in Z$.

The first assumption is straightforward: a valid measure Y_{ij} should contain information about the latent variable beyond pure noise. The second assumption requires that the treatment Z_i has at least two distinct values and is randomly assigned to units, implying that it is independent of the potential latent outcomes. We further assume that each outcome's measurement error ϵ_{ji} is independent of the treatment – as Stoetzer et al. (2022) point out, this assumption amounts to an exclusion restriction because it implies that the treatment does not affect the errors we make when measuring the latent outcomes. The fourth assumption stipulates that the measurement errors are unrelated to the latent potential outcomes. This assumption implies, for example, that higher values of the latent potential outcomes are no more likely to coincide with higher values of the measurement errors than lower values of the latent potential outcomes. The SUTVA assumption is a standard requirement for any study of cause and effect; in this context, it implies that potential outcomes remain stable regardless of which subjects receive treatment, implying, for example, no spillovers between subjects.

Because the latent variable η_i has no inherent scale, without loss of generality, we let $\lambda_1 = 1$ so that we can interpret η using the same metric as Y_1. For example, if η_i represents a latent distance, and Y_{i1} is scaled in terms of kilometers while Y_{2i} is scaled in terms of miles, setting $\lambda_1 = 1$ means that effects on η_i are scaled in terms of kilometers (Bollen, 1989, pp.239-240). (The choice of which observed outcome to use for scaling purposes is arbitrary and has no effect on the statistical

significance of the estimated ATE.[3]) The advantage of this "unstandardized" scaling approach over approaches that standardize all outcome measures to have unit variance (or scale the latent outcome to have unit variance) is that standardization imposes scaling parameters that are sample-specific. As Loehlin (1998, pp.28-29) points out, standardization can lead a researcher to mistakenly conclude that two estimated ATEs are different when they are in fact the same. Consider two very large studies in which the intervention exerts the exactly same average treatment effect on a latent outcome measured by the same Y_{ij} indicators. Standardization would produce two quite different estimates of the ATE if the measurement error variances were much larger in the first study than the second.[4] Furthermore, ad hoc standardization approaches are ill-suited to parameter estimation. In their assessment of several commonly-used methods of creating standardized indexes as experimental outcomes — simple indexes, weighted indexes, and principal components factor scores — Blair et al. (2023, section 15.4) conclude that *none* of them recovers a known parameter in repeated simulations.[5] In short, maintaining interpretable scaling units rather than standardizing simplifies the identification problem and preserves the comparability of results across experimental replications.

This setup presupposes that the relationship between the latent variable η_i and the observed measurements Y_{ij} is linear. We would offer two comments about linearity. First, linearity implicitly requires that all measures are valid in the sense that they all measure the latent variable η_i up to the scale factor λ_j. Formally, we can see this linear modeling assumption from the expression $Y_{ij} = \lambda_j Y_{i1} + \epsilon_{ij} \quad \forall j \neq 1$. As we will show later, we can assess the linearity assumption empirically, and in the Kalla and Broockman (2020) application below, the measures indeed exhibit the expected linear relationship. Second, when outcomes are binary or ordinal, we can still use a linear model to obtain a linear approximation; the problem, as we note below, is that assessing how well the model fits the data becomes more model-dependent. For this reason, our design recommendations focus on gathering outcomes in ways that satisfy the linearity assumption rather than shoehorn discrete outcome measures into a linear modeling framework. The formal results and corresponding strategies are discussed in the section 6.2.

3 Identification of Measurement Parameters

We seek to identify the average causal effect of Z_i on the latent variable η_i. However, η_i is not directly observed, and we can only partially measure it given the unknown scaling parameters λ_j and unobserved measurement errors ϵ_{ij}. If we knew λ_j, we could rescale the observed measures, for example, by $\frac{Y_{ij}}{\lambda_j} = \eta_i + \epsilon'_{ij}$, to approximate the latent variable η_i using the same units as Y_{i1}. The question is how to identify λ_j. As demonstrated in the following proposition, identification requires a third variable to serve as an instrumental variable. What is perhaps surprising is that this

[3]Beware of the fact that SEM in Stata uses its estimated standard errors to calculate p-values, but these standard errors are calculated based on numerical derivatives that are subject to scale-specific error. To obtain scale invariant p-values, conduct a likelihood ratio test comparing the log-likelihood of the fitted model to the log-likelihood of the restricted model, which constrains the average treatment effect to be zero.

[4]Indeed, the standardized estimates would differ in this example even though a simple regression of Y_{i1} on Z_i in each sample would, in expectation, generate identical estimates.

[5]In the SI A, we show that both of the estimators we propose below successfully recover the ATE in their simulations (up to scale factor, depending on which outcome measure is used to set the metric). Moreover, the estimated standard error from SEM closely matches the empirical standard error from their simulation.

third variable can be either the treatment Z_i or other measurements Y_{ij} under the relatively mild conditions in Assumption 1.[6]

Proposition 1 (Identification of the measurement scaling parameters)**.** *Suppose Assumption 1 holds. Then in the above causal framework given $\lambda_1 = 1$, λ_j is identified either by*

(1) $\lambda_j = \frac{Cov(Z_i, Y_{ij})}{Cov(Z_i, Y_{i1})}$ *if* $\mathbb{E}[\eta_i^1 - \eta_i^0] \neq 0$ *or by*

(2) $\lambda_j = \frac{Cov(Y_{ik}, Y_{ij})}{Cov(Y_{ik}, Y_{i1})} \forall k \neq 1$ *and* $k \neq j$*, if* $Cov(\epsilon_{ik}, \epsilon_{i1}) = Cov(\epsilon_{ij}, \epsilon_{ik}) = 0$ *and* $Var[\eta_i] \neq 0$.

Proof. The proof is in the SI. □

Proposition 1 suggests that λ_j can be identified via an instrumental variables approach. We can use either the treatment Z_i or other outcome measures Y_{ij} as instrumental variables when certain assumptions hold. Specifically, Z_i must be a "relevant" instrumental variable whose average treatment effect on η_i is nonzero (so that the denominator of the ratio converges to a nonzero value as the sample size increases). The exclusion assumption of Z_i as an instrument follows from Assumption 1B, because it is independent of the measurement errors. Similarly, Y_{ik} is a valid instrumental variable if the measurement errors are uncorrelated with one another, and the limiting covariance between Y_{ij} and Y_{i1} is nonzero. Even when the measurement errors are correlated, Z_i remains a valid instrument so long as the treatment is unrelated to these errors of measurement.

To illustrate the reasoning behind the instrumental variables approach, consider one strategy for identifying λ_2. Recall that $Y_{i2} = \lambda_2(\eta_i^0 + Z_i\tau_i) + \epsilon_{i2}$. This equality implies the following relationship between Y_{i2} and Y_{i1}

$$Y_{i2} = \lambda_2 Y_{i1} + (\epsilon_{i2} - \lambda_2 \epsilon_{i1})$$

Because Y_{i1} is correlated with the error ϵ_{i1}, λ_2 is not directly identified, and we cannot simply regress Y_{2i} on η_i because the latter is unobserved. Again, instrumental variables regression provides a consistent estimator. Given the excludability of the randomly assigned treatment variable Z_i (Assumption 1C), we can use Z_i as an instrumental variable for Y_{i1}. The (super population) covariance between Z_i and Y_{i2} is

$$Cov(Z_i, Y_{i2}) = \lambda_2 Cov(Z_i, Y_{i1}) + Cov(Z_i, \epsilon_{i2} - \lambda_2 \epsilon_{i1})$$

In the proof, we show that if the average treatment effect is nonzero ($\mathbb{E}\tau_i \neq 0$), then $Cov(Z_i, Y_{i1}) \neq 0$. In that case, Z_i is a "relevant" predictor of Y_{i1}. Moreover, given that Z_i is randomly assigned and assumed to be independent of the measurement errors, Z_i is a valid instrumental variable for Y_{i1}. Therefore, as n increases, $\frac{\widehat{Cov}(Z_i, Y_{i2})}{\widehat{Cov}(Z_i, Y_{i1})}$ converges to λ_2.

The same approach shows that Y_{i3} can also serve as an instrumental variable for Y_{i1}. Suppose n were very large. The limiting covariance between Y_{i1} and Y_{i2} is

$$Cov(Y_{i2}, Y_{i3}) = \lambda_2 Cov(Y_{i3}, Y_{i1}) + Cov(Y_{i3}, \epsilon_{i2} - \lambda_2 \epsilon_{i1})$$

Given that $Cov(Y_{i3}, Y_{i1}) \neq 0$ (which is implied by conditions in proposition 1) and measurement errors are independent of the latent traits and the errors of measurement, ($Cov(\epsilon_{i3}, \epsilon_{i1}) =$

[6] When pre-treatment covariates are available, they can also serve as instrumental variables, as discussed in Section 6.1.

$Cov(\epsilon_{i3}, \epsilon_{i2}) = 0$). Y_{i3} is also a valid instrument; therefore, a consistent estimator is $\hat{\lambda}_2 = \frac{\widehat{Cov}(Y_{i2},Y_{i3})}{\widehat{Cov}(Y_{i1},Y_{i3})}$. The sample covariances that we observe empirically provide plug-in estimators. Conveniently, an equivalent plug-in approach may be implemented using instrumental variables regression.[7]

What about cases such as Figure 1, where the experiment includes *both* a randomly assigned treatment and three measures of the latent variable? In such cases, we have more than one plug-in estimator for λ_2 and λ_3, and therefore the unknown scaling parameters are said to be overidentified. Overidentification is helpful in two ways. First, we can combine the plug-in estimators to form a more efficient estimator, using method of moments or maximum likelihood. The former has the advantage of making no distributional assumptions about the observed or unobserved variables in model, while the latter imposes the assumption of multivariate normality. In practice, the two estimators tend to produce similar estimates in large samples when the observed variables are distributed symmetrically (Olsson et al., 2000; Browne, 1984; Yuan and Chan, 2005; Boomsma and Hoogland, 2001).

The second advantage of overidentification is that it allows us to test whether the data accord with the posited model. If two or more plug-in estimators render markedly different estimates, that is a sign that at least one modeling assumption is untenable. For example, when Z_i is used as the instrumental variable for identifying λ_2, we need not invoke any assumptions about the independence of measurement errors (i.e., the assumption that ϵ_{i1}, ϵ_{i2}, and ϵ_{i3} are independent of one another), whereas when we identify λ_2 using Y_{i3} as the instrumental variable, we do invoke this assumption when asserting that the limiting covariances among the errors are zero. If the two IV estimates differ markedly, that is a sign that independent measurement errors may be an untenable assumption.

Maximum likelihood estimation makes it straightforward to evaluate the extent to which an observed set of covariances deviates from the values implied by the estimated parameters. A conventional goodness-of-fit statistic allows a researcher to test the null hypothesis that the stipulated model (under multivariate normality) generated the data (Bollen, 1989). Other things being equal, the larger the number of randomized treatments and outcome measures, the more powerful such hypothesis tests become.

4 Estimation of Average Treatment Effects

Once the λ_j are identified, researchers have many options for estimating the average treatment effect on the latent variable. These options fall into two broad categories. The first category involves forming a weighted average of the observed outcome measures to approximate the latent variable, albeit with some residual error. This index-building approach has the advantage of allowing the analyst to use conventional methods, such as regression, to estimate the ATE. The complication that arises with two-step estimation of the ATE is estimation of standard errors, since the uncertainty associated with estimating the weights will not be incorporated into conventional regression standard errors. An alternative approach is to estimate both the measurement parameters and the ATE simultaneously as part of a system of linear equations. This approach is typically referred to as structural equation modeling (SEM). Although SEM is often associated with full-information

[7] See SI D.

estimators that assume multivariate normality (Jöreskog, 1970), the same models can be rooted in more agnostic assumptions and estimated by method-of-moments (Kline, 2023).

This section discusses these two estimation approaches – weighted indices and maximum likelihood – using simulated data to assess their ability to recover known treatment effects and estimate sampling variability. Fortunately, both approaches tend to perform well and generate similar results. Two-step methods have some advantages in terms of transparency, as they allow the analyst to display regression results visually (Cook, 2009); maximum likelihood has the advantage of providing a unified framework for estimating treatment effects and accompanying standard errors. Both methods are suitable for hypothesis testing using randomization inference.

4.1 Difference-in-means based on a Weighted Scaled Index

The first estimator is the weighted scaled difference-in-means estimator. For each individual i, we first construct a new outcome measure by creating a weighted average of the various outcome measures. Taking this weighted average to be the outcome, we use difference-in-means to estimate the causal effect. Formally, for each individual i, let $\tilde{Y}_{ij} = \frac{1}{\lambda_k} Y_{ij}$, and $\tilde{Y}_i = \sum_{j=1}^{J} \omega_j \tilde{Y}_{ij}$, where ω_j is the weight and $\sum_{j=1}^{k} \omega_j = 1$. A naive choice of the weight is $\frac{1}{k}$. That is, the new outcome measure $\tilde{Y}_i$ is an average of all measurements for individual i. Although simple averages are widely used in practice (e.g., Ansolabehere et al., 2008), they are not optimal from the standpoint of estimating the ATE as precisely as possible. A more efficient approach is to use inverse-variance weighting. Because we divide each measure by λ_j, the true variance of the measurement error is $\frac{\sigma^2(\epsilon_{\cdot j})}{\lambda_j^2}$; this leads to the optimal weight $\omega_j^* = \frac{\lambda_j^2/\sigma^2(\epsilon_{\cdot j})}{\sum_{j=1}^{k} \lambda_j^2/\sigma^2(\epsilon_{\cdot j})}$, and the corresponding variance given each individual is $\frac{1}{\sum_{j=1}^{J} \lambda_j^2/\sigma^2(\epsilon_{\cdot j})}$. This optimal weighting scheme assumes that measurement errors are independent; when errors are correlated, the optimal weighting algorithm becomes more complex.[8]

Using the weighted outcome measure as a proxy for the latent outcome, a researcher applies conventional methods for estimating the ATE. The most straightforward way to estimate the ATE using the proxy measure is the difference-in-means estimator:

$$\begin{aligned}\hat{\tau} &= \frac{1}{n_1}\sum_{i=1}^{n} Z_i \tilde{Y}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i \\ &= \frac{1}{n_1}\sum_{i=1}^{n} Z_i [\sum_{j=1}^{k} \omega_j \tilde{Y}_{ij}] - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)[\sum_{j=1}^{k} \omega_j \tilde{Y}_{ij}]\end{aligned}$$

Suppose we know the true λ_j. Then, it is straightforward to see that the oracle estimator $\hat{\tau}^*$ is unbiased. When λ_j is unknown, we can still use a consistent estimator of λ to obtain a consistent difference-in-means estimator.

[8] In general, the weight is derived by minimizing the variance of weighted average $\sum_{j=1}^{J} \omega_j Y_{ij}$, subject to the constraint that $1'w = 1$. Define Σ as the variance-covariance matrix of the measurement errors. By using the Lagrange multiplier method, we get the weight $w = (1'\Sigma^{-1}1)^{-1}(1'\Sigma)$. If measurement errors are independent, the optimal weight reduces to $\omega_j^* = \frac{\lambda_j^2/\sigma^2(\epsilon_{\cdot j})}{\sum_{j=1}^{k} \lambda_j^2/\sigma^2(\epsilon_{\cdot j})}$.

Proposition 2. *Suppose assumption 1 holds. Consider a weighted difference-in-means estimator* $\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i$.

(1) If λ_j *is known, then* $\hat{\tau}$ *is unbiased.*

(2) If λ_j *is estimated consistently under proposition 1, then the weighted difference-in-means estimator is asymptotically unbiased.*

Proof. The proof is in the SI. □

To work out the variance of this estimator, we start with the simplest case in which the λ_j scaling factors are known based on the measurement properties observed in prior studies. Under this scenario, we can treat each λ_j as a constant and ignore its uncertainty, in which case the variance is similar to the Neyman variance. We define the weighted outcome with respect to the potential η as follows: $\tilde{Y}_i^1 := \sum_{j=1}^{k} \frac{\omega_j\lambda_j}{\hat{\lambda}_j}[\eta_i^1 + \epsilon_{ij}]$ and $\tilde{Y}_i^0 := \sum_{j=1}^{k} \frac{\omega_j\lambda_j}{\hat{\lambda}_j}[\eta_i^0 + \epsilon_{ij}]$. The population variance for two treatment groups is $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}[\tilde{Y}_i^1 - \overline{\tilde{Y}}_i^1]^2$ and $S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}[\tilde{Y}_i^0 - \overline{\tilde{Y}}_i^0]^2$. Their unbiased sample analogue is $\hat{S}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} Z_i(\tilde{Y}_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i)^2$ and $\hat{S}_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(1-Z_i)(\tilde{Y}_i - \frac{1}{n_1}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i)^2$. Also, let $\Delta\tilde{y}_i := \tilde{Y}_i^1 - \tilde{Y}_i^0 = \frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1 - \eta_i^0)$ be the variance of the individual treatment effect.

Proposition 3. *Suppose assumption 1 holds. Consider the variance estimator* $\widehat{Var}_Z(\hat{\tau}) = \frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_0^2}{n_0}$.

(1) In the finite population framework $N < \infty$*, the variance of the weighted difference-in-means estimator is* $Var_Z(\hat{\tau}) = \frac{\mathbb{E}_\epsilon[S_1^2]}{n_1} + \frac{\mathbb{E}_\epsilon[S_0^2]}{n_0} - \frac{\mathbb{E}_\epsilon[S_{\Delta\tilde{y}}^2]}{n}$*. The variance estimator* $\widehat{Var}_Z(\hat{\tau})$ *is conservative.*

(2) In the super-population framework $N = \infty$*, the variance of the weighted average estimator is* $\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$*, where* $\sigma_1^2 = Var_{sp}[\tilde{Y}_i^1] = \mathbb{E}_{sp}[(\tilde{Y}_i^1 - \mathbb{E}_{sp}[\tilde{Y}_i^1])^2]$ *and* $\sigma_0^2 = Var_{sp}[\tilde{Y}_i^0]$*. The variance estimator* $\widehat{Var}_Z(\hat{\tau})$ *is unbiased.*

Proof. The proof is in the SI. □

In the finite population framework, the estimator $\widehat{Var}_Z(\hat{\tau}) = \frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_0^2}{n_0}$ is conservative because the part we cannot estimate $\frac{\sigma_{\Delta\tilde{y}}^2}{n}$ is nonnegative; this conclusion is similar to the traditional Neyman variance in the finite population framework. Suppose instead that observations are randomly drawn from a super-population. By analogy to a completely randomized experiment a fully observable outcome, the third part that cannot be estimated in the variance of the oracle estimator (i.e. the variance of individual latent treatment effects: $-\frac{\sigma_{\Delta\tilde{y}}^2}{n}$) disappears in the super-population framework. Therefore, $\widehat{Var}_Z(\hat{\tau})$ is unbiased.

If the λ_j are estimated from the experimental data, we have to account for the estimation uncertainty. (Fortunately, this uncertainty tends to be relatively small. See SI C.) Defensible estimates of the standard error of the estimated ATE, accounting for the estimation of the measurement parameters, may be obtained using GMM, as discussed in SI E. The Pseudo-algorithm is summarized in SI F.

4.2 Stacking Estimator

In practice, regression is widely used to estimate the ATE, especially when covariates are used to improve precision. It is straightforward to show that the above difference-in-means estimator is equivalent to the "stacked" regression described below.

Using rescaled outcomes, stacked regression estimates the ATE for all outcome measures in a single pass. Let $[\tilde{Y}_i] = (\tilde{Y}_{i1}, \tilde{Y}_{i2}, ..., \tilde{Y}_{iJ})^T$ denote the stacked re-scaled outcomes for individual i; and let $[\tilde{Y}] = ([\tilde{Y}_1], [\tilde{Y}_2], ..., [\tilde{Y}_n])^T$ be the $nJ \times 1$ outcome variable vector. We also stack treatment assignment so that $[Z_i] = (Z_i, Z_i, ..., Z_i)^T$ is the $J \times 1$ vector and $[Z] = ([Z_1], [Z_2], ..., [Z_n])^T$. Then we define $X = \begin{bmatrix} 1, 1, ..., 1 \\ [Z_1], [Z_2], ..., [Z_n] \end{bmatrix}^T$ as the $nJ \times 2$ matrix.

Consider the stacked regression where $[\tilde{Y}]$ is regressed on X. Then the stacking estimator $\hat{\beta}$ is the OLS coefficient for the regressor $[Z]$.

Proposition 4. *In the stacking regression where $[\tilde{Y}]$ is regressed on X. The OLS estimator of the coefficient of X is equivalent to* $\frac{1}{n_1} Z_i \sum_{i=1}^{n} (\frac{1}{k} \sum_{j=1}^{J} \tilde{Y}_{ij}) - \frac{1}{n_0} (1 - Z_i) \sum_{i=1}^{n} (\frac{1}{k} \sum_{j=1}^{J} \tilde{Y}_{ij})$.

Proof. The proof is in the SI. □

Recall the weighted difference-in-means estimator is $\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} Z_i [\sum_{j=1}^{J} \omega_j \tilde{Y}_{ij}] - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) [\sum_{j=1}^{J} \omega_j \tilde{Y}_{ij}]$. We can see the stacking regression estimator is equivalent to the weighted difference-in-means estimator when $\omega_j = \frac{1}{k}$. To achieve an optimal weighting scheme, we can simply multiply the optimal weight ω_j^* by $\tilde{Y}_{ij}$ before stacking, i.e. $[\tilde{Y}_i] = (\omega_1^* \tilde{Y}_{i1}, ..., \omega_J^* \tilde{Y}_{iJ})^T$. If the researcher seeks to adjust for covariates, they can simply be added to the regression specification. Covariate-adjusted stacked regression provides a consistent estimator of the average treatment effect on the latent outcome, and its variance may be smaller than the unadjusted stacked regression when the covariates are prognostic of the observed outcomes.

4.3 Structural Equation Modeling

An alternative approach is to model all of the parameters in Figure 1 as a system of linear equations, one equation for each of the endogenous variables in the causal graph. Figure 1 implies one equation for the latent outcome variable (η_i) and three equations for each of the observed outcome measures. After imposing the scaling metric $\lambda_1 = 1$, we are left with a total of 8 free parameters: the variance of the randomized treatment (ϕ), the average treatment effect of Z_i on η_i, the variance of the unobserved causes of the latent outcome variable (ψ), two scaling parameters (λ_2 and λ_3), and three variances of the measurement errors associated with Y_{i1}, Y_{i2}, Y_{i3}. These 8 parameters may be used to predict the 10 observed variances and covariances among the four measured variables (Z_i,Y_{i1},Y_{i2},Y_{i3}). Since some of these parameters are overidentified, the estimator we use may affect our empirical estimates because different estimators use different objective functions when gauging the fit between the estimated parameters and the observed variance-covariance matrix. As noted above, the most commonly used estimator is maximum likelihood under the assumption of multivariate normality, but other estimators that make weaker distributional assumptions are available as well. MLE and GLS both tend to perform well when the number of observations is large relative to the number of parameters (Yuan and Chan, 2005).

5 Variance Trade-off: More Outcome Measures or More Subjects?

One implication of the preceding section is that, as Broockman et al. (2017, p.449) argue, researchers could improve the precision of their experiments by investing in additional outcome measures. Under what conditions is such an investment warranted?

Recall that the super-population variance of the (oracle) estimator is

$$Var_{sp}(\hat{\tau}) = \frac{Var_{sp}[\tilde{Y}_i^1]}{n_1} + \frac{Var_{sp}[\tilde{Y}_i^0]}{n_0} \tag{1}$$

This formula shows that researchers can decrease this variance by adding more measurements, which decreases the variance in the numerator, or by adding more observations, which increases the denominator. We examine both approaches and explore the optimal allocation under a budget constraint.

5.1 Advantages of Adding More Outcome Measurements

The variances of the difference-in-means or regression estimators depend mainly on the variance of the weighted averages $\tilde{Y}_i^1$ and $\tilde{Y}_i^0$. Assuming $Cov(\epsilon_j, \epsilon_k) = 0$ and ignoring the estimation variance of $\hat{\lambda}$ on the grounds that it is known based on existing studies, under optimal inverse-variance weighting ω_j^*, the variance of $\tilde{Y}_i^1$ is $\sigma^2(\eta_i^1) + \frac{1}{\sum_{j=1}^J \lambda_j^2/\sigma^2(\epsilon_{\cdot j})}$, where the latter part is due to (optimally weighted) measurement error.

Given J existing measurements, a researcher may consider including additional outcome measures. To explore the potential payoff of doing so, we simplify the notation by using Σ^J to denote the denominator $\sum_{j=1}^J \lambda_j^2/\sigma^2(\epsilon_{\cdot j})$. The change in variance of $Var_{sp}(\tilde{Y}_i^1)$ resulting from the addition of one more measurement may be expressed as follows:

$$\frac{1}{n_1}\frac{1}{\Sigma^J + \frac{\lambda_{J+1}^2}{\sigma^2(\epsilon_{\cdot J+1})}} - \frac{1}{n_0}\frac{1}{\Sigma^J}$$

This equation shows that the inclusion of a highly reliable outcome measure (i.e., a measure with a high ratio of trait variance to error variance) may substantially increase the precision with which the ATE is estimated. However, the marginal gains from each additional measure depend on the reliability of the measures already contained in the additive index. Fortunately, with optimal weighting, in expectation more measurements never increase the variance of the estimated ATE because unreliable measures are down-weighted accordingly.[9]

[9]It is important to note that this desirable feature of adding outcome measures may not hold without optimal weighting. For example, with equal weights, the variance of $\tilde{Y}_i^1$ is $\frac{1}{n_1}\sigma^2(\eta_i^1) + \frac{1}{n_0}\frac{1}{J^2}\sum_{j=1}^J \frac{\sigma^2(\epsilon_{\cdot j})}{\lambda_j^2}$. The gain from adding one more measure is

$$\frac{1}{(J+1)^2}\frac{\sigma^2(\epsilon_{\cdot J+1})}{\lambda_{J+1}^2} - (\frac{1}{J^2} - \frac{1}{(J+1)^2})\sum_{j=1}^J \frac{\sigma^2(\epsilon_{\cdot j})}{\lambda_j^2}$$

Suppose the new outcome measure is quite noisy: the variance of the measurement error $\sigma^2(\epsilon_{\cdot J+1})$ is large. Therefore,

A simple scenario illustrates the gains from including an additional measure and applying optimal weights. For ease of exposition, we assume a balanced design ($n_0 = n_1$) and consider the case in which all outcome measures have the same error variance $\sigma^2(\epsilon)$ and factor loading λ. Therefore, $\Sigma_J = J\Sigma$, where $\Sigma := \frac{\lambda^2}{\sigma^2(\epsilon)}$. In this case, where all measures are equally reliable, the inclusion of an additional measure from current J measures can be expected to change the variance of the estimated ATE by

$$\Delta(J) := \frac{4}{n}\frac{-1}{J(J+1)\Sigma}$$

Due to the rapidly increasing quadratic term $J(J+1)$ in the denominator, the payoff diminishes with each additional measurement. Moreover, the precision gain from the $(J+1)^{th}$ measure depends on Σ, inverse of the variance of measurement error, or the precision of the measure. As the measurement error variance increases, the precision improvement from an additional measure increases.

We can get a better feel for the advantages of including an additional outcome measure by examining the percentage change. We compare the reduction in variance from adding the $(J+2)^{th}$ measure with the reduction in variance from adding the $(J+1)^{th}$ measure.

$$\frac{\Delta(J+1)}{\Delta(J)} = \frac{J}{(J+2)}$$

For example, suppose that we have only one outcome measure, $J = 1$. The denominator is the variance reduction by adding the second measure. The numerator is the variance reduction of the third measure. The gain is $\frac{1}{3} \approx 33.33\%$. If we add a fourth measure, the variance reduction will be $\frac{1}{6} = 16.67\%$ compared to the gain from the second measure.[10] From this example, we surmise that when outcome measures are equally reliable, three or four measures are sufficient in practice, although more measures may be helpful for other reasons, such as linearizing the relationship between η and sub-indices of outcome measures.

5.2 Gathering More Observations

In the super-population variance of our optimal weighted average estimator shown in Equation (1), the numerators are the variances of weighted average potential outcomes in the super-population, which are fixed. Therefore, adding more observations only affects the denominator. Under a balanced design, when one adds j more observation in each group, the variance will be $[\frac{2Var[\tilde{Y}_i^1]}{n+2j} + \frac{2Var[\tilde{Y}_i^0]}{n+2j}]/[\frac{2Var[\tilde{Y}_i^1]}{n} + \frac{2Var[\tilde{Y}_i^0]}{n}] = \frac{n}{n+2j}$ of the variance when the sample size is n; in other words, the percentage reduction in variance is $\frac{2j}{n+2j}$. Therefore, researchers can add $\frac{n}{4}$ subjects to the treatment and control groups to decrease the variance by $\frac{1}{3}$.

the first term can dominate the second term in the formula, which makes the variance increase rather than decrease. For this reason, those who advocate the creation of simple additive indexes caution that there is no guarantee that including an additional measure will improve reliability on the margin (Ansolabehere et al., 2008, p.218).

[10] Simulation results illustrating these diminishing returns are shown in the SI K.

5.3 Trade-off under a budget constraint

We can apply this framework to a more general budget allocation problem. Suppose that researchers have a budget $B > 0$ and the cost of adding one more measure is c_1 and the cost of adding one more observation is c_2. Consider the optimal allocation of budget for the sample size n and the number of items k. This is a typical constrained optimization problem where the objective function is the variance $\frac{2[\sigma^2(\eta_i^1)+\sigma^2(\eta_i^0)]}{n} + \frac{4}{nJ\Sigma}$. The objective is to minimize the variance given the budget constraint $c_1 J + c_2 n \leq B$. Optimization may be used to find the best budget allocation between additional outcome measures and additional respondents.

Now, let us consider a more concrete scenario. Suppose researchers have an extra $\$5000$ budget. They may decide to recruit more respondents and/or add more measures. What is the optimal decision? To be consistent with the previous simulation, we assume the current sample size is $n = 500$, and there is only one measure. The marginal cost of the measure is $c_1 = \$1000$ and the marginal cost of each respondent is $c_2 = \$10$. Let us consider two scenarios: the outcomes are measured with high reliability (0.75) or low reliability (0.4).[11]

For high reliability case, the optimal solution is to add one more measure and recruit 400 more respondents. However, for the low reliability case, the optimal solution is to add two more measures and recruit 300 more respondents. In other words, although the extra measures are low quality, they are an especially good investment. This somewhat surprising result can be explained by our previous calculation for $\Delta(J) := \frac{4}{n}\frac{-1}{J(J+1)\Sigma}$ and the simulation in the figure A.10. When measurement error is large (low reliability), the variance reduction is relatively greater.

Two other considerations also arise when weighing the advantages of collecting additional outcome measures. The first is that additional observable manifestations of the latent outcome make for more overidentifying restrictions on the key scaling parameters. Each additional measure gives us more ways to assess the scaling properties of all of the outcome measures at hand. This fact, in turn, allows us to assess the goodness-of-fit of a posited multi-equation model. The second consideration is that the more outcome measures one gathers, the more likely one is to discover the inadequacies of one or more of the measures. One symptom of a measurement problem is poor fit between the predicted and actual variance-covariance matrix; another is fluctuation of the estimated ATE when certain outcome measures are included or excluded. Ideally, additional outcome measures improve precision and robustness; however, the inclusion of flawed measures may undercut these benefits. Therefore, when gauging the trade-off between allocating resources to subjects or outcome measurements, one should bear in mind that the value of additional outcome measures hinges on their validity and reliability.

[11] To be specific, we set $Var(\eta^1) = Var(\eta^0) = 1$. For low reliability, the variance of the measurement error is 1.701 so that $\frac{Var(\eta)}{Var(Y)} \approx 0.4$. For high reliability, the variance of the measurement error is 0.379 so that $\frac{Var(\eta)}{Var(Y)} \approx 0.75$. The objective function is to maximize the variance gain, or equivalently, $\min_{k,j} \frac{-2k[Var(\eta^1)+Var(\eta^0)]}{500(500+k)} + \frac{-4(k+500j+jk)}{500(1+j)\Sigma(500+k)}$, where j is the additional measures and even number k is the additional respondents. The constraint is $1000j + 10k \leq 5000$.

6 Strategies for Improving Precision and Assessing Robustness

6.1 Covariates

The analysis of experiments with latent outcomes may benefit in three ways from the inclusion of covariates, defined as variables measured prior to random assignment. First, covariates may be used to identify measurement parameters, such as the λ_k in Figure 1. Indeed, when it comes to identifying scaling parameters, covariates play the same identifying role as randomly assigned interventions, and the assumptions required are similar: the covariate must be statistically independent of the measurement errors but correlated with the structural disturbance term (ζ_i). In other words, the covariate must predict Y_{i1} to some extent so that the instrumental variables estimator is defined. Second, covariates that are strongly prognostic of outcomes will improve the precision with which the average treatment effects are estimated. This property of covariate adjustment is in keeping with conventional experimental analysis, where outcomes are modeled directly without reference to latent variables. Third, because covariates contribute to the identification of measurement parameters, covariates allow the researcher to use overidentification to assess the goodness-of-fit of the posited multi-equation model. In sum, although covariates are not strictly necessary for the identification or estimation of measurement parameters, they nevertheless may play a useful role, especially when they are predictive of outcomes.

6.2 Measurement Strategies to Satisfy the Linearity Assumption

One key assumption of our approach is the linear relationship between η_i and the measured outcomes Y_{ik}. This assumption is clearly untenable when survey outcomes are measured with binary response options. One work-around is to use more granular response options, such as presenting an agree-disagree question with response options ranging from "strongly agree" to "strongly disagree." However, this approach may still yield a lopsided response distribution that would not plausibly be modeled as a linear function of the latent variable.

A more convincing way to satisfy the linearity condition is to create an additive index composed of several discrete measures, in much the same way that the Scholastic Aptitude Test measures verbal ability by counting the number of correct answers to many specific multiple-choice questions.[12] This approach, which dates back to the early days of psychometric assessment, may require pilot testing (perhaps in a nonexperimental context or during pre-treatment baseline measurement) in order to devise a series of discrete measures that, when summed, produce granular distributions with few observations in each tail.

Intuitively, we can think of each (discrete) measurement Y_{ij} as being drawn from some distribution indexed by a latent variable η_i. More generally, suppose that the expectation of this distribution is given by $L_j(\eta_i, \alpha_j)$, where L is a linear transformation, and α_j can be a vector of parameters (item difficulty for example). Then, if the correlations between measures are not too strong, the sum of these measurements will converge to $\sum_{i=1}^{J} L_j(\eta_i, \alpha_j)$. As a result, we would expect the additive index to have a linear relationship with η_i. This formulation encompasses many common

[12] Modern variants of scholastic aptitude tests adaptively calibrate the difficulty of the questions based on the answers to prior questions, and the same cost-saving principle can be applied to survey measures of political knowledge and other topics (Montgomery and Cutler, 2013).

data-generating processes, for example, binary measurements (e.g., agree/disagree) drawn from a Bernoulli distribution or ordinal responses from an ordered probit model, among others.

Proposition 5. *For each i, let $\{Y_{ij}\}$ be random variables with finite means $L_j(\eta_i, \alpha_j) < \infty$ and $\sup_j Var(Y_{ij}) < \infty$. Define $\overline{Y}_i = \frac{1}{J}\sum_{i=1}^{J} Y_{ij}$.*

If $\sum_{J=1}^{\infty} \frac{\sum_{j=1}^{J}\sum_{k=1}^{J} Cov(Y_{ij},Y_{ik})}{J^2} < \infty$, then $\overline{Y}_i$ converges to a linear function of η_i, given that $\lim_{J\to\infty} \frac{1}{J}\sum_{j=1}^{J} L_j(\eta_i, \alpha_j)$ exists.

Proof. The proof is in the SI. □

If the measures for each individual were independent conditional on their latent score, then the above results reduce to Kolmogorov's Strong law of large numbers. The IRT modeling suggested by Stoetzer et al. (2022) generally satisfies the above proposition as well. In sum, our framework offers a more flexible method to estimate latent treatment effects without assuming constant treatment effects or untestable nonlinear modeling assumptions.

To illustrate how additive indices help satisfy the assumed linear mapping between latent and observed outcomes, we performed a series of simulations. In each simulation, binary responses were generated using an IRT model, and two outcome indices were then created by summing half of the binary items together.[13] The number of items used to create each index increases from one simulation to the next, as indicated by the horizontal axis labels of Figure 2. The sequence of graph panes shows that as we sum more binary items, the LOESS fitted line – which allows a nonlinear relationship if the data suggest it – between the observed measures and the true η_i used in the simulation becomes increasingly linear. Figure 2 illustrates an extreme case where all component items are binary. In SI G, we also show the simulation results for ordinal items. Intuitively, if the component items have more granular response options, fewer items will be needed to form an additive index that meets the linearity condition.[14] On the other hand, building an index from items that are subject to the same measurement errors due to question wording, order, and format slows the rate at which linearity is achieved. As a thought experiment, consider the limiting case in which two measures generate exactly the same responses; in this case, an index based on both measures is no more granular than each measure considered separately. The design implication is clear: if possible, researchers should measure the latent variable in different ways so as to reduce the correlation between errors of measurement (Green et al., 1993).

6.3 Testing Measurement Equivalence between Treatment and Control Groups

Our basic framework assumes that responses from subjects in the treatment and control groups have the same measurement structure. In other words, the translation from η_i to the Y_{ik} is the same regardless of the subject's experimental condition. This assumption may not hold in practice, for example, if there is something about the treatment that changes the way that subjects interpret outcome questions.[15] Different scaling parameters jeopardize the exclusion restriction,

[13] We use simTrt function in the psych R package to generate the data. To be specific, the discrimination parameter $a \sim Uniform(.5, 1.5)$, item difficulties $b \sim N(0, 2)$, and the guessing asymptote is $c = 0.2$.

[14] In practice, η is not observed. Instead, we could show the increasingly linear properties between two observed measures. Figure A.2, which does so, conveys the same point: as the number of binary items used to create an index increases, the more linear the apparent relationship.

[15] The intuition extends to non-survey outcomes as well. If outcomes in the control group are measured in miles but outcomes in the treatment group are measured in kilometers, any apparent mean difference may be misleading.

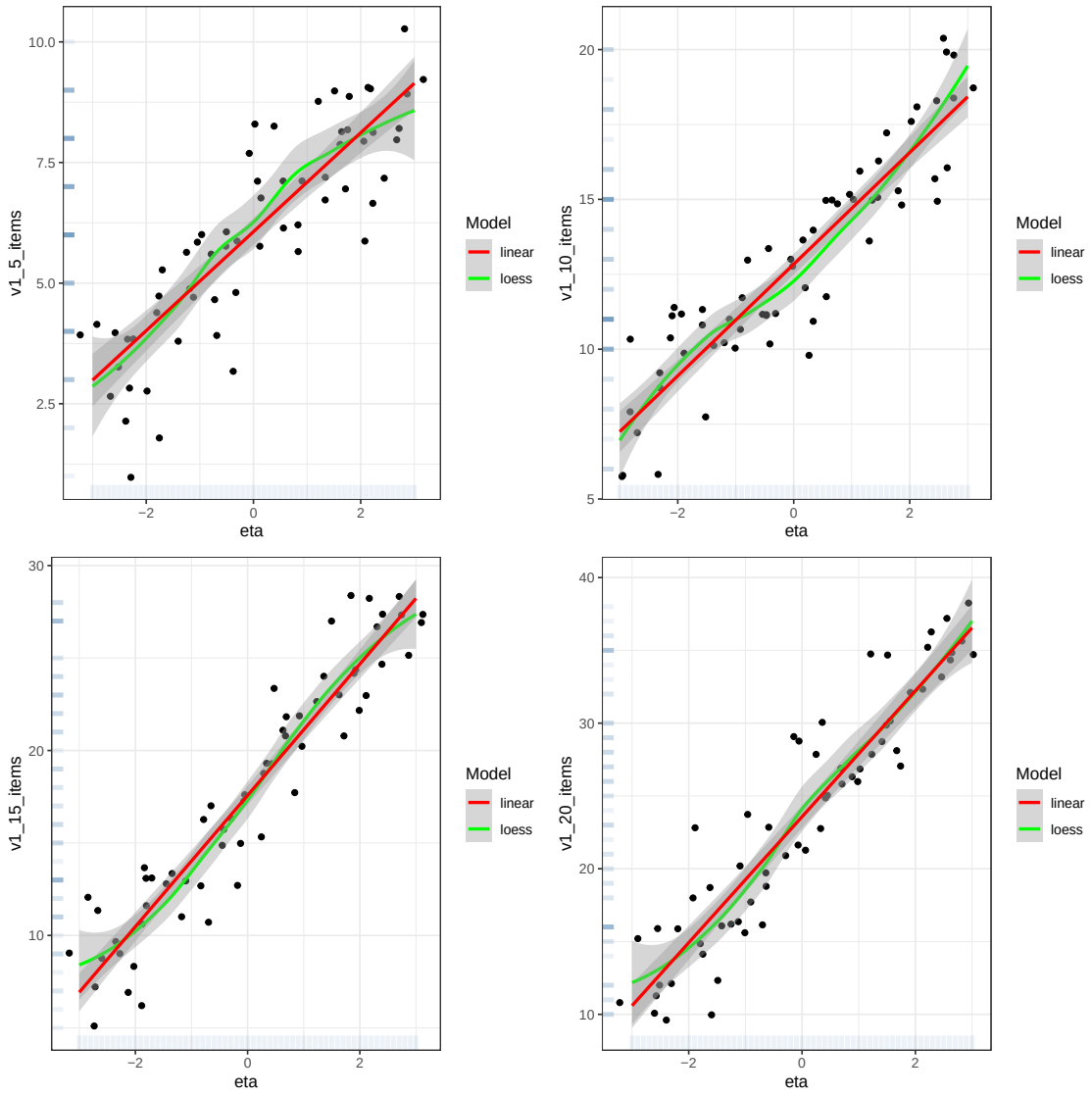

Figure 2: **Linearity between η and an additive index created by adding binary items together**. We create an additive index v_1 by summing up 5, 10, 15, and 20 binary responses from IRT models. Data points have been jittered slightly for clarity. The rug plots on both axes denote the distribution of data. The vertical axis shows the additive index created by adding all binary variables in the simulation. The horizontal axis is the true latent variable (η) used in the IRT model.

which holds that the treatment itself is the only systematic reason that expected outcomes in the treatment and control groups may differ. Fortunately, the measurement equivalence assumption can be assessed empirically. In order to make the test as general as possible, we relax the implicit assumption that the observed outcome is centered on the latent variable. A more flexible model is that $Y_{ij} = \lambda_j g(\eta_i) + \alpha_j + \epsilon_{ij}$, where α_j denotes the so-called "structural mean" (Bagozzi and Yi, 1989) for experimental group j. Unlike the model presented above, this one expresses the treatment effect as a shift in latent means, while at the same time allowing for the possibility that the measurement scaling parameters λ_j and $VAR(\epsilon_{ij})$ may differ for the treatment and control groups.

When we have at least two outcome measures and at least one covariate, the measurement scaling parameters can be identified separately for each experimental group. Suppose we label the treatment group's scaling parameters λ'_j and the control group's scaling parameters λ''_j. Under the hypothesis of measurement equivalence, $\lambda'_j = \lambda''_j$, which is testable using conventional tests based on normal theory or more agnostic tests rooted in randomization inference. An analogous null hypothesis may be imposed on the variances of the measurement errors (i.e., $VAR(\epsilon'_j) = VAR(\epsilon''_j)$, and the constraints on the scaling factors and measurement error variances may be tested jointly. In sum, the structural means model is a nested alternative to a model that imposes the same measurement structure to observed outcomes in both the treatment and control groups. Comparing the two models empirically can provide some assurance that the more restrictive model, which imposes an exclusion restriction, is consistent with the data.

6.4 Comparison to Seemingly Unrelated Regression (SUR)

Many experiments are designed with a clear intention to gather multiple measures of a given latent outcome, in which case the modeling framework described here is an appropriate starting point. However, other experiments gather multiple outcomes that are not necessarily viewed as redundant measures of a given latent factor. For example, field experiments on conditional cash transfers examine outcomes such as whether children in the household are enrolled in school, vaccinated, and free from signs of malnutrition. One could view such outcomes as manifestations of an underlying factor (child well-being), but one could instead consider them to be three distinct outcomes. How might we evaluate the adequacy of the two modeling approaches?

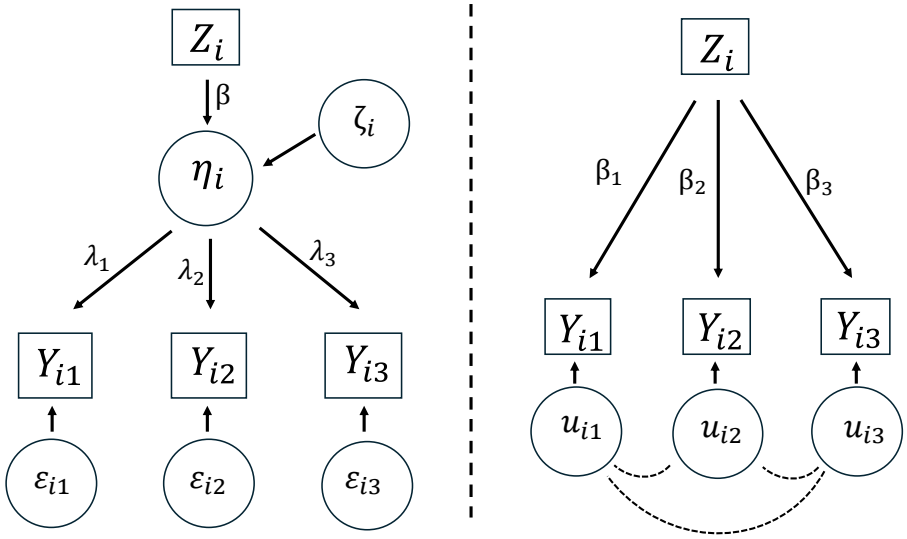

Figure 3: Graphical Depiction of a Linear Measurement Process with Additive Measurement Errors (left) and a Seemingly Unrelated Regression Model (right)

Fortunately, the two models may be expressed as nested alternatives. Figure 3 shows two causal graphs side-by-side. The left pane depicts the now-familiar model in which three outcome measures are linear manifestations of a latent outcome. The right pane depicts the seemingly unrelated

regression model (Zellner, 1962) in which the same three outcome measures are modeled as distinct outcomes.[16] Whereas the latent variable model has 8 free parameters, the SUR model has 10 because it does not constrain the causal path from Z_i to the Y_{ik} to flow through η_i. Therefore, two degrees of freedom may be used to assess the extent to which the more parsimonious latent variable model adequately fits the observed variance-covariance matrix. Rejection of the null hypothesis calls into question one or more assumptions of the latent variable model in favor of the more agnostic SUR model. Conversely, failure to reject the null hypothesis suggests that the causal pathways implied by the latent variable model are approximately the same as the three distinct average treatment effects depicted in the SUR model (i.e., $\beta\lambda_k \approx \beta_k$). On theoretical grounds, the latent variable model may not be applicable to many experiments that measure multiple outcomes. But when the latent variable model is plausible for a given application, the fact that its parameterization has testable implications means that researchers need not rely solely on theoretical intuition when choosing between the latent variable model and the SUR model.

7 Application

Like Stoetzer et al. (2022), we draw our empirical example from the Kalla and Broockman (2020) field experiment designed to assess the effects of two different door-to-door canvassing interventions on subjects' views concerning immigrants. Kalla and Broockman's Study 1 shows that canvassing conversations that feature a "non-judgmental exchange of narratives" produce persistent changes in subjects' attitudes about immigrants, where as otherwise similar conversations that omit this persuasive strategy seem to produce weaker effects. One attractive feature of this study is the fact that it devoted considerable attention and resources to outcome measurement. As summarized in Table A.1, five questions gauged respondents' attitudes towards undocumented immigrants. A separate battery of three questions measured respondents' views about policy proposals concerning expediting paths to citizenship and reducing threats of deportation.

The two treatment conditions (full treatment Z_1 and abbreviated treatment Z_2) in conjunction with two primary outcome scales (attitudes Y_1 and policy views Y_2) give us four observed variables and therefore ten observed variances and covariances. If we model these data by assuming that the attitude index and the policy views index both measure a common underlying factor, the two parameters of central interest are the average treatment effect of full treatment on the latent outcome and the average treatment effect of abbreviated treatment on the latent outcome. The latent variable model is depicted in the left pane of figure 4.

Let us review and evaluate the assumptions under which both causal parameters in the latent variable model are identified: the treatments are randomly assigned and therefore independent of latent potential outcomes; the treatments affect measured outcomes only insofar as they influence the latent outcome (the exclusion restriction, which implies that the measurement errors for each observed outcome (ϵ_{ki}) are independent of the interventions); the stable unit treatment value

[16]The weighted index method proposed by Anderson (2008) is closely connected to the SUR model because it does not posit a latent variable measured with error. This method down-weights outcomes that are correlated with one another in order to generate a unified outcome measure with minimum variance. Simulations in the Figure A.4 show that Anderson's method is less powerful than SEM when outcomes measures truly tap a latent outcome with error. On the other hand, when the data generating process is a SUR model, Anderson's method is slightly more efficient than SUR when the disturbances are highly correlated.

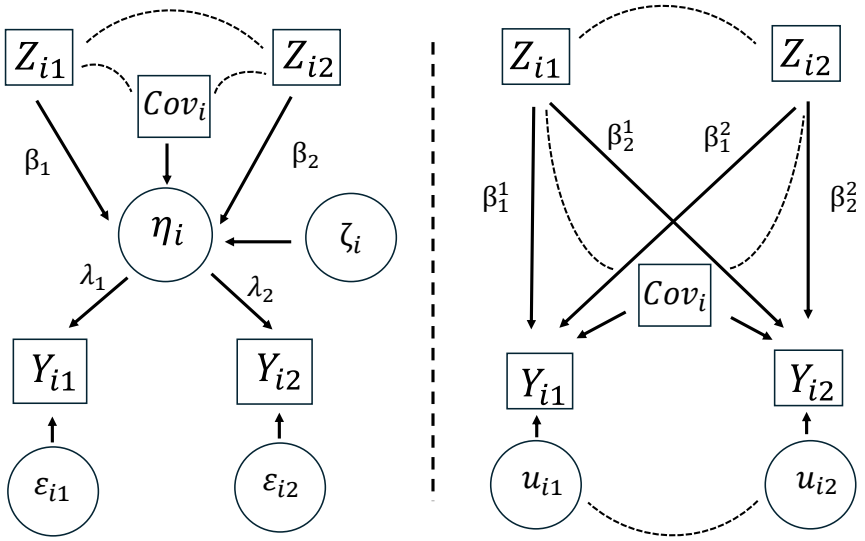

Figure 4: Graphical Depiction of Kalla and Broockman (2020) with Additive Measurement Errors (left) and a Seemingly Unrelated Regression Model (right).

assumption (SUTVA), which implies no spillovers between subjects; at least one of the two treatments has a nonzero ATE on the latent outcome; and observed outcomes are linear functions of the latent outcome. Next, let us evaluate the plausibility of each assumption.

In this application, random assignment is satisfied by the experimental design and remains tenable in the wake of attrition, which seems to be unrelated to treatment assignment. The exclusion restriction implies that the interventions do not affect the outcome measures except insofar as they affect the underlying factor (Stoetzer et al., 2022). This assumption could be jeopardized if subjects drew the connection between the treatment they received and the survey, but this study used an unobtrusive measurement strategy that did not alert participants to the connection between the series of surveys they completed and their incidental visit from a canvasser. SUTVA violations due to spillover effects seem unlikely to have a material effect on outcomes here given the separation among subjects from different households. The requirement that the interventions have a nonzero effect on the latent outcome only comes into play in this application because we have limited ourselves to just two outcome measures and have excluded covariates; as we will see below, we find evidence that at least one treatment influenced the latent outcome, and covariate adjustment leaves the basic pattern of results unchanged.

As for the linearity assumption, one of the study's strengths is its use of multi-item outcome measures. These measures gauge outcomes with sufficient granularity that we can visually inspect the scatterplot between Y_1 and Y_2. As shown in A.8, fitting a flexible LOESS curve through this plot reveals a near-linear pattern that coincides with the fitted least-squares regression line. A linear mapping from the latent outcome to each of the observed outcome measures seems plausible in this application.[17]

With these identifying assumptions, consistent estimation is straightforward. If we declare the scale of the latent outcome variable to be in the same units as the attitude index (i.e., $\lambda_1 = 1$,

[17]Tables A.2 and A.4 may be used to calculate the proportion of observed variance in the outcomes that is attributable to the latent variable η_i. For attitudes, this proportion is 0.74, and for policy views it is 0.83.

the ATE of the full treatment on the latent factor can be estimated consistently using the approach described in Section 4.[18] Estimation of the ATE of the abbreviated treatment is consistently estimated in an analogous fashion. The fact that we have more information than we need to identify certain parameters such as λ_2 means that can (1) compare alternative estimates of each parameter and (2) leverage the overidentifying information to obtain more precise estimates. Comparing alternative estimates is one way to assess model fit; when two or more estimates of the same parameter diverge, the implication is that one or more underlying assumptions are incompatible with the data. When leveraging overidentifying assumptions, the most commonly used estimation approach assumes that the random components of the model are drawn independently from normal distributions and estimates the parameters using maximum likelihood estimation (Bollen, 1989). More agnostic approaches allow for departures from normality (Browne, 1984).

Estimates of the two interventions' average treatment effects on the latent outcome are presented in Table 1. The first two columns present estimates obtained using maximum likelihood, and columns (3) and (4) report regressions in which an optimally weighted average of the two outcomes are regressed on the two treatments.[19] Columns (5) and (6) report SEM estimates, this time using a GLS estimator rather than MLE and using bootstrapped standard errors. Each estimator is presented using two specifications, one without covariate adjustment and a one adjusting for respondents' baseline attitudes about immigrants (See SI J). Comparing columns (1) and (3) shows that MLE and regression generate almost identical estimates. The regression estimator appears to produce smaller standard errors, but this advantage is illusory; the two-step method underestimates the standard errors because no account is taken of the uncertainty associated with scaling parameter that is used to generate the weighted outcome measure. Comparing columns (2) and (4) again shows that the two estimators render similar estimates. Covariate adjustment greatly improves precision and clearly indicates that the full treatment was effective in changing subjects' opinions. The MLE covariate-adjusted estimates also suggest that the full treatment was more effective than the abbreviated treatment, with a differential effect of 0.341 (SE=0.141). The MLE estimates and standard errors are almost identical to those obtained using GLS and bootstrapped standard errors.

Next, we compare the results from Table 1 to estimates obtained using the seemingly unrelated regression model, which makes no assumptions about why the two outcomes may be correlated. Table 2 presents two different SUR specifications, with and without covariate adjustment. The first column, without covariate adjustment, reports two separate OLS regressions in which attitudes (Y_1) and policy views (Y_2) are regressed on the two treatment indicators. The second column repeats these two regressions, this time including baseline covariates. Although the estimated treatment effects look different in Table 1 and Table 2, some simple manipulations show that the two tables tell very similar stories. The estimated effects on Y_{i1} are scaled identically in the two tables, and the two estimates are quite close. The estimated effects on Y_{i2}, however, differ because the latent variable model applies that scaling factor $\widehat{\lambda}_2 = 1.653$ when estimating the ATE in column (1) of Table 1. If we divide the SUR estimate by this estimate of λ_2, we come very close to the estimate reported in Table 1. The same holds for all of the coefficients reported in the two tables, implying that the modeling constraints of the latent variable model fit are quite compatible with the agnostic

[18] Because this application features two randomly assigned interventions (Z_{i1} and Z_{i2}) that are correlated, consistent estimates are obtained using multiple regression rather than bivariate regression.

[19] When creating the weighted average, we estimate the scaling parameter using method-of-moments. The optimal weight $\omega_j^* = \frac{\hat{\lambda}_j^2/\hat{\sigma}^2(\epsilon_{\cdot j})}{\sum_{j=1}^{k} \hat{\lambda}_j^2/\hat{\sigma}^2(\epsilon_{\cdot j})}$, where the variance of measurement error is calculated from MLE.

Table 1: Estimation Results: MLE, Weighted Index, and GLS

	MLE	MLE	Index	Index	GLS	GLS
	(1)	(2)	(3)	(4)	(5)	(6)
treat (full)	0.384*	0.431***	0.384*	0.430***	0.384*	0.431***
	(0.213)	(0.099)	(0.196)	(0.099)	(0.213)	(0.096)
treat (mod)	0.224	0.090	0.225	0.090	0.224	0.090
	(0.206)	(0.101)	(0.200)	(0.101)	(0.215)	(0.102)
baseline covariate		0.662***		0.662***		0.662***
		(0.011)		(0.009)		(0.012)
λ_2	1.653	1.549	1.652	1.549	1.653	1.549
χ^2 $p-value$	0.923	0.978			0.923	0.978
Adjusted R^2			0.001	0.762		
Degree of freedom	1	2				

Note: Index denotes Difference-in-means based on an optimal weighted index. Sample size $N = 1,578$. λ_2 in columns 1, 2, 5, and 6 are estimated by SEM; λ_2 in columns 3 and 4 are estimated by 2SLS using the two treatments and, where applicable, covariates as instrumental variables. For GLS, the standard errors are computed from 10,000 bootstraps. The MLE models are estimated by R package lavaan, and the GLS models are estimated by R package systemfit. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

SUR model, which by construction fits the observed variance-covariance matrix perfectly. A more formal test of the fit of the latent variable model against the fit of the SUR model comes to the same conclusion: the p-value of the likelihood-ratio test is 0.980, so we have no basis to reject the adequacy of the (more parsimonious) latent variable model.

Table 2: Estimation Results: Seemingly Unrelated Regression

	Model 1	Model 2
eq1: treat_full	0.383	0.418***
	(0.214)	(0.120)
eq1: treat_mod	0.231	0.088
	(0.218)	(0.122)
eq2: treat_full	0.635	0.689***
	(0.335)	(0.192)
eq2: treat_mod	0.364	0.143
	(0.342)	(0.196)
eq1: cov_bseline		0.662***
		(0.011)
eq2: cov_bseline		1.025***
		(0.018)
eq1: R^2	0.002	0.688
eq2: R^2	0.002	0.673
eq1: Adj. R^2	0.001	0.687
eq2: Adj. R^2	0.001	0.673

Note: In the model 1, two equations represent two regression models: eq1 is attitudes $\sim$ treat (full) + treat (mod) and eq2 is policy views $\sim$ treat (full) + treat (mod). In the model 2, we add baseline covariates. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Our final specification check is to assess whether the measurement properties of the outcome measures operate symmetrically for all three experimental groups. The null hypothesis is that the measures operate in the same way across the three groups; the alternative model is that the measurement parameters – the scaling factors and the measurement error variances – differ across groups. A total of 6 degrees of freedom differentiate the two models. The p-value of the model comparison is 0.171, suggesting that there may be some concern about differential measurement across groups, but the gap in terms of goodness-of-fit is not decisive.[20]

[20]In their analysis of differential item functioning when applying a hierarchical IRT model to these data, Stoetzer et al. (2022) found similarly equivocal results.

8 Conclusion

This paper has shown how experiments with redundant outcome measures may be analyzed using latent outcome models while retaining the agnosticism of the design-based framework. In contrast to Stoetzer et al. (2022), we have assumed a linear relationship between the latent outcome and the observed outcomes, which enabled us to draw connections to other widely used tools such as instrumental variables regression. Whether linearity is a plausible assumption depends on what is being measured and how; crucially, linearity may be addressed at the design stage of an experiment and assessed empirically. For example, outcome measures may be constructed from batteries of survey questions, and the resulting indices may be sufficiently granular to warrant a linear modeling framework. In such cases, identification and estimation of causal effects becomes straightforward. In other words, an investment in outcome measurement pays dividends when it comes time to analyze the experimental results with less reliance on restrictive parametric assumptions.

Another advantage of working within a linear modeling framework is that it allows the researcher to make use of statistical tools that are routinely used in experimental analysis, ordinary least-squares regression and instrumental variables regression. Instrumental variables regression may be used to estimate the scaling parameters that link the latent outcome to the observed outcomes, which in turn may be used to estimate the error variances associated with each observed outcome measure. With estimates of these scaling parameters and error variances in hand, a researcher may easily create an optimally weighted average of the observed outcomes that approximates the latent outcome and has an interpretable scale. This weighted average may be regressed on the treatments and covariates in the usual way, and the relationship between interventions and the latent outcome are easily visualized using scatter plots. Alternatively, structural equation models may be used to estimate the scaling parameters, error variances, and causal parameters in a single statistical procedure. Both approaches produce consistent estimates of the average treatment effect on the latent outcome. The latter approach, although less familiar to many researchers and less amenable to data visualization, has the advantage of estimating standard errors more defensibly.

Another important message of this paper is the value of overidentification via additional outcome measures, treatments, or covariates. These layers of redundancy both facilitate robust estimation of scaling parameters and create opportunities for goodness-of-fit tests that can help assess the empirical adequacy of the posited latent variable model. Failure to pass such tests implies that future studies need to invest in new or improved outcome measures. When outcomes are measured via surveys, the iterative process of proposing and evaluating measures may involve changes to question wording or response options, or something more fundamental, such as ensuring that respondents are attentive (and human).

Latent variable models are optional but potentially helpful. When an outcome cannot be observed directly, researchers should strive to design experiments that measure this latent outcome in multiple ways. For decades, statisticians have pointed out that one way to increase the power of an experiment is to measure the outcome more reliably. The algebra in section 5 reiterates this point, showing the gains in precision that occur with each additional measure of a latent outcome. When maximizing the precision with which the average treatment effect on the latent outcome is estimated, it sometimes makes more sense to invest in additional outcome measures rather than additional subjects.

This paper has focused on the challenge of measuring a single latent outcome, but the framework may be expanded to designs in which two or more latent outcomes are posited. (See the SI L

for an example.) The development of valid and reliable measures of multiple latent constructs has a long history in psychometric research (Anderson and Gerbing, 1988; Campbell and Fiske, 1959). The details of this type of investigation go beyond the scope of this paper but typically involve an iterative process of proposing measures that seem to tap into a theoretically defined construct (face validity), properly distinguish subjects that are known to differ on the latent dimension of interest (construct validity), and show the expected patterns of high and low correlations with measures of other constructs (convergent and discriminant validity) (Chan, 2014). The systematic study of measurement involves a combination of theoretical and empirical steps that may be conducted outside the confines of an experiment but potentially provides valuable insights for experimental applications.

References

Alwin, D. F. and Tessler, R. C. (1974). Causal models, unobserved variables, and experimental data. *American Journal of Sociology*, 80(1):58–86.

Anderson, J. C. and Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3):411.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansolabehere, S., Rodden, J., and Snyder, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102(2):215–232.

Bagozzi, R. P. (1977). Structural equation models in experimental research. *Journal of Marketing Research*, 14(2):209–226.

Bagozzi, R. P. and Yi, Y. (1989). On the use of structural equation models in experimental designs. *Journal of Marketing Research*, 26(3):271–284.

Blair, G., Coppock, A., and Humphreys, M. (2023). *Research design in the social sciences: declaration, diagnosis, and redesign*. Princeton University Press.

Bollen, K. A. (1989). *Structural equations with latent variables*, volume 210. John Wiley & Sons.

Boomsma, A. and Hoogland, J. J. (2001). The robustness of lisrel modeling revisited. *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog*, 2(3):139–168.

Broockman, D. E., Kalla, J. L., and Sekhon, J. S. (2017). The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4):435–464.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1):62–83.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81.

Chan, E. K. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In *Validity and validation in social, behavioral, and health sciences*, pages 9–24. Springer.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons.

Costner, H. L. (1971). Utilizing causal models to discover flaws in experiments. *Sociometry*, pages 398–410.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. W.W. Norton.

Green, D. P., Goldman, S. L., and Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of personality and social psychology*, 64(6):1029.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

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

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2):i–41.

Kalla, J. L. and Broockman, D. E. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2):410–425.

Kano, Y. (2001). Structural equation modeling for experimental data. *Structural equation modeling: Present and future*, pages 381–402.

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

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Lawrence Erlbaum Associates Publishers.

Montgomery, J. M. and Cutler, J. (2013). Computerized adaptive testing for public opinion surveys. *Political Analysis*, 21(2):172–192.

Olsson, U. H., Foss, T., Troye, S. V., and Howell, R. D. (2000). The performance of ml, gls, and wls estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural equation modeling*, 7(4):557–595.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Sävje, F., Aronow, P., and Hudgens, M. (2021). Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2):673.

Sorbom, D. (1981). Structural equation models with structured means. *Systems under indirect observation*, pages 183–195.

Stoetzer, L. F., Zhou, X., and Steenbergen, M. (2022). Causal inference with latent outcomes. *American Journal of Political Science*.

Su, F., Mou, W., Ding, P., and Wainwright, M. J. (2023). A decorrelation method for general regression adjustment in randomized experiments. *arXiv preprint arXiv:2311.10076*.

Yuan, K.-H. and Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika*, 70(4):791–798.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Supplementary Information

A Declare Design

In their discussion of multiple outcomes, Blair et al. (2023, section 15.4), consider several index-creation methods for outcome measurement.[21] Their simulation model posits an experimental intervention on a latent outcome, the same problem that we consider in this paper. However, none of their index-creation methods – additive standardized indexes or predicted values based on factor scores – successfully recovers the target ATE from their simulation. In this section, we apply our two proposed estimators to their setup: the optimal-weighted scaled index and the SEM (maximum likelihood) estimator.

In their simulation, the latent variable is generated by the equation $\eta = 1 + X + 2 * rnorm(N)$. The true ATE of the experimental intervention is therefore 1. The three observed outcome measures are generated as follows: $Y_1 = 3 + 0.1 * \eta + rnorm(N, sd = 5)$,$Y_2 = 2 + 1 * \eta + rnorm(N, sd = 2)$, and $Y_1 = 1 + 0.5 * \eta + rnorm(N, sd = 1)$. Notice that the three scaling factors (λ_j) are $\{0.1, 1, 0.5\}$, so if we set $\lambda_2 = 1$, we will get outcomes in the original metric. If instead we set $\lambda_1 = 1$, the model fit will be identical, but now the scale in which the ATE is measured will be changed by a factor of 10. Which outcome measure we use to set the scale is arbitrary (e.g., millimeters vs. centimeters), but we must remember to interpret the ATE in terms of whatever scale we select.

Figure A.1 shows the empirical distribution of our two estimators across 1000 simulations, each with a sample of 500 subjects. For ease of interpretation, we set $\lambda_2 = 1$ so that the target ATE parameter remains 1.0. Both of these estimators produce estimates that are centered around the true average treatment effect. Further examination of the empirical sampling distributions shows that both approaches work well, with SEM doing a better job of estimating the true standard error.

B Proof

B.1 Proof of Proposition 1

Proof. Note that $Y_{ij} = \lambda_j Y_{i1} + (\epsilon_{ij} - \lambda_j \epsilon_{i1})$.

We show (1) first. The formula suggests that Z can be a valid IV. The population covariance ($n \to \infty$) between Z_i and Y_{ij} is

$$Cov(Z_i, Y_{ij}) = \lambda_2 Cov(Z_i, Y_{i1}) + Cov(Z_i, \epsilon_{ij} - \lambda_2 \epsilon_{i1})$$

[21] This example is drawn from code in section 15.4 of the book, which focuses on estimating the conditional mean of a latent variable. We build on the authors' data-generating process to study the related question of estimating an average treatment effect of an experimental intervention using what is effectively a difference-in-means estimator. Note that we could have reparameterized the SEM model as a structured means model in order to estimate the intercepts as well.

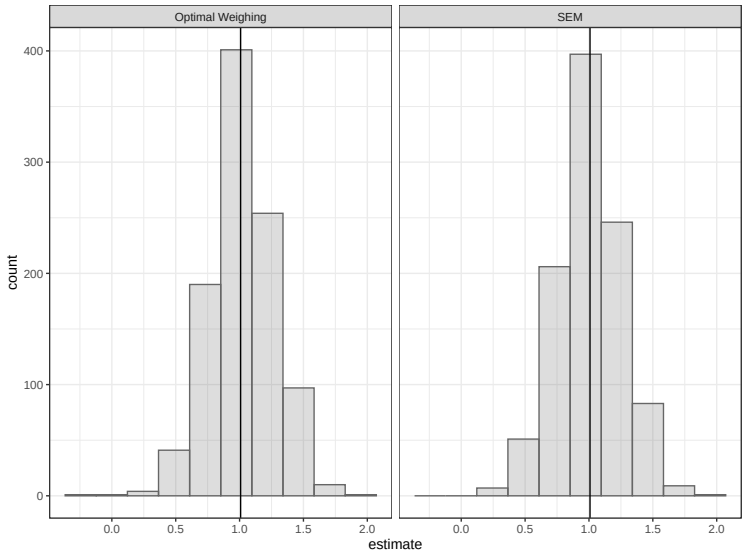

Figure A.1: **Declare Design Simulation**. Empirical distribution of optimal-weighted scaled index and SEM estimators. The vertical line is the target ATE of the intervention on the latent outcome, which is scaled to equal to one in the simulation.

Assumption of $\mathbb{E}[\eta_i^1 - \eta_i^0] \neq 0$ implies that $Cov(Z_i, Y_{i1}) \neq 0$. To see this, $Cov(Z_i, \eta_i^0 + \tau_i Z_i + \epsilon_{i1}) = Cov(Z_i, \eta_i^0) + \mathbb{E}\tau_i[\mathbb{E}Z_i^2 - \tau_\square[\mathbb{E}Z_i]^2] = \mathbb{E}\tau_i Var(Z_i)$. Moreover, under assumption 1, Z_i is independent of measurement errors. Therefore, Z satisfies relevance and exclusion assumptions of IV. The IV estimator $\lambda_j = \frac{Cov(Z_i, Y_{ij})}{Cov(Z_i, Y_{i1})}$ follows immediately.

For (2), still starting from the equation $Y_{ij} = \lambda_j Y_{i1} + (\epsilon_{ij} - \lambda_j \epsilon_{i1})$, we observe that the covariance between Y_{ij} and Y_{ik} $(k \neq 1, k \neq j)$ is

$$Cov(Y_{ik}, Y_{ij}) = \lambda_2 Cov(Y_{ik}, Y_{i1}) + Cov(Y_{ik}, \epsilon_{ij} - \lambda_j \epsilon_{i1})$$

We check $Cov(Y_{ik}, Y_{i1})$ first. Under assumption 1,

$$\begin{aligned} Cov(Y_{ik}, Y_{i1}) &= Cov[\lambda_k(\eta_i^0 + \tau_i Z_i), \eta_i^0 + \tau_i Z_i] + \lambda_k Cov(\epsilon_{ik}, \epsilon_{i1}) \\ &= \lambda_k Var[\eta_i^0 + \tau_i Z_i] + 0 \\ &= \lambda_k Var[\eta_i] \end{aligned}$$

Then, $Cov(Y_{ik}, Y_{i1}) \neq 0$ if $Var[\eta_i] \neq 0$. With assumptions that $Cov(\epsilon_{ik}, \epsilon_{i1}) = Cov(\epsilon_{ik}, \epsilon_{ij}) = 0$, Y_{ij} also satisfies exclusion assumption of IV. Therefore, the IV estimator $\lambda_j = \frac{Cov(Y_{ik}, Y_{ij})}{Cov(Y_{ik}, Y_{i1})}$ is consistent.

□

B.2 Proof of Proposition 2

Proof. Consider (1) first.

Note that $\mathbb{E}[\tilde{Y}_i^1] = \eta_i^1$ and $\mathbb{E}[\tilde{Y}_i^0] = \eta_i^0$.

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^*] &= \mathbb{E}\{\frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i\} \\
&= \mathbb{E}\{\frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i^0\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}\eta_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}\frac{n_0}{n}\eta_i^0 \\
&= \frac{1}{n}\sum_{i=1}^{n}\eta_i^1 - \frac{1}{n}\sum_{i=1}^{n}\eta_i^0 \\
&= \tau
\end{aligned}$$

For (2),

$$\mathbb{E}[Z_i\tilde{Y}_i] = \mathbb{E}[Z_i(\sum_{j=1}^{k}\omega_j\tilde{Y}_{ij})] \tag{2}$$

$$= \mathbb{E}\{Z_i[\sum_{j=1}^{k}\omega_j\frac{\lambda_j}{\hat{\lambda}_j}(Z_i\eta_i^1 + (1-Z_i)\eta_i^0 + \frac{1}{\lambda}\epsilon_i^j)]\} \tag{3}$$

$$= \mathbb{E}[Z_i^2\sum_{j=1}^{k}\omega_j\frac{\lambda_j}{\hat{\lambda}_j}\eta_i^1] + \mathbb{E}[Z_i(1-Z_i)\sum_{j=1}^{k}\omega_j\frac{\lambda_j}{\hat{\lambda}_j}\eta_i^0] \tag{4}$$

$$= \mathbb{E}[Z_i]\eta_i^1[\sum_{j=1}^{k}\omega_j\lambda_j\mathbb{E}(\frac{1}{\hat{\lambda}_j})] \tag{5}$$

$$= \mathbb{E}[Z_i]\eta_i^1[\sum_{j=1}^{k}\omega_j\lambda_j\mathbb{E}(\frac{1}{\lambda_j} - \frac{1}{\lambda_j^2}(\hat{\lambda}_j - \lambda_j) + o_p(\hat{\lambda}_j - \lambda_j))] \tag{6}$$

$$= \frac{n_1}{n}\eta_i^1 + [-\mathbb{E}[Z_i]\eta_i^1\sum_{j=1}^{k}\omega_j\frac{1}{\lambda_j}\mathbb{E}[\hat{\lambda}_j - \lambda_j + o_p(\hat{\lambda}_j - \lambda_j)]] \tag{7}$$

where (4) follows $\mathbb{E}\epsilon_i^j = 0$, and (6) follows that $\mathbb{E}[Z_i^2] = \mathbb{E}[Z_i]$ and $\mathbb{E}[Z_i(1-Z_i)] = 0$.

Take the limit, and all bias terms in the squared bracket go to 0. Therefore, $\lim_{n\to\infty}\mathbb{E}[Z_i\tilde{Y}_i] = \frac{n_1}{n}\eta_i^1$

Similarly, we have $\lim_{n\to\infty}\mathbb{E}[(1-Z_i)\tilde{Y}_i] = \frac{n_0}{n}\eta_i^0$.

Therefore, we can show that the weighted difference-in-means estimator $\hat{\tau}$ is asymptotically

unbiased.

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[\hat{\tau}] &= \lim_{n\to\infty}\mathbb{E}\{\frac{1}{n_1}\sum_{i=1}^{n}Z_i\tilde{Y}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i\}\\ &= \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n}\eta_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}\frac{n_0}{n}\eta_i^0\\ &= \frac{1}{n}\sum_{i=1}^{n}\eta_i^1 - \frac{1}{n}\sum_{i=1}^{n}\eta_i^0\\ &= \tau\end{aligned}$$

□

B.3 Proof of Proposition 3

Proof. To calculate the variance of this estimator, we apply the rule of total variance,

$$Var(\hat{\tau}) = \mathbb{E}_\epsilon[Var_Z(\hat{\tau}|\epsilon)] + Var_\epsilon[\mathbb{E}_Z[\hat{\tau}|\epsilon]]$$

First to calculate the $Var_Z(\hat{\tau}|\epsilon)$. Recall our weighted average estimator

$$\begin{aligned}\hat{\tau} &= \frac{1}{n_1}\sum_{i=1}^{n}Z_i\tilde{Y}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i\\ &= \frac{1}{n_1}\sum_{i=1}^{n}Z_i\tilde{Y}_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i^0\\ &= \frac{1}{n_1}\sum_{i=1}^{n}Z_i[\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1+\epsilon'_{ij})] - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)[\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^0+\epsilon'_{ij})]\end{aligned}$$

Because we fix ϵ and $\hat{\lambda}_j$, $\tilde{Y}_i = \sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i+\epsilon'_{ij})$ is constant. Therefore, the variance is exactly the Neyman's variance in the finite population.

$$\begin{aligned}Var_Z(\hat{\tau}|\epsilon) &= Var_Z\{\sum_{i=1}^{n}Z_i[\frac{\tilde{Y}_i^1}{n_1}+\frac{\tilde{Y}_i^0}{n_0}] - \frac{1}{n_0}\sum_{i=1}^{n}\tilde{Y}_i^0\}\\ &= \frac{n_1n_0}{n(n-1)}\sum_{i=1}^{n}\{\frac{\tilde{Y}_i^1}{n_1}+\frac{\tilde{Y}_i^0}{n_0}-\frac{\overline{\tilde{Y}}_i^1}{n_1}-\frac{\overline{\tilde{Y}}_i^0}{n_0}\}^2\\ &= \frac{S_1^2}{n_1}+\frac{S_0^2}{n_0}-\frac{S_{\Delta\tilde{y}}^2}{n}\end{aligned}$$

where $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}[\tilde{Y}_i^1 - \overline{\tilde{Y}}_i^1]^2$, $\Delta\tilde{y}_i := \tilde{Y}_i^1 - \tilde{Y}_i^0 = \frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1 - \eta_i^0)$, and $S_{\Delta\tilde{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}[\Delta y_i - \overline{\Delta\tilde{y}}]^2$, which is the variance of the weighted individual treatment effects.

Next, we calculate $\mathbb{E}_Z[\hat{\tau}|\epsilon]$.

$$\begin{aligned}\mathbb{E}_Z[\hat{\tau}|\epsilon] &= \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i^1 - \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i^0 \\ &= \frac{1}{n}\sum_{i=1}^{n}[\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1 - \eta_i^0)]\end{aligned}$$

And thus the variance $Var_\epsilon[\mathbb{E}_Z[\hat{\tau}|\epsilon]] = 0$.

Therefore, the total variance is

$$\begin{aligned}Var_Z(\hat{\tau}) &= \mathbb{E}_\epsilon[\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_{\Delta\tilde{y}}^2}{n}] + Var_\epsilon[\mathbb{E}_Z[\hat{\tau}|\epsilon]] \\ &= \frac{\mathbb{E}_\epsilon[S_1^2]}{n_1} + \frac{\mathbb{E}_\epsilon[S_0^2]}{n_0} - \frac{\mathbb{E}_\epsilon[S_{\Delta\tilde{y}}^2]}{n} + 0 \\ &= \frac{\mathbb{E}_\epsilon[S_1^2]}{n_1} + \frac{\mathbb{E}_\epsilon[S_0^2]}{n_0} - \frac{\mathbb{E}_\epsilon[S_{\Delta\tilde{y}}^2]}{n}\end{aligned}$$

These expectation terms can be explicitly calculated. For example,

$$\begin{aligned}\mathbb{E}[S_1^2] &= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}[\tilde{Y}_i^1 - \overline{\tilde{Y}}_i^1]^2 \\ &= \frac{1}{n-1}\mathbb{E}[\sum_{i=1}^{n}(\tilde{Y}_i^1)^2 - n(\overline{\tilde{Y}}_i^1)^2] \\ &= \frac{1}{n-1}\mathbb{E}\{\sum_{i=1}^{n}[\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1 + \epsilon'_{ij})]^2 - n[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(\eta_i^1 + \epsilon'_{ij})]^2\}\end{aligned}$$

We can then take expectations of each term. The expectation of terms with a single ϵ' is 0; the expectation of terms with ϵ'^2 is $\mathbb{E}[\epsilon'^2_{ik}] = Var[\epsilon'_{ik}] = \frac{1}{\lambda_k^2}\sigma^2(\epsilon_{ik})$; and the expectation of $\mathbb{E}[\epsilon'_{ij}\epsilon'_{ik}] = \frac{1}{\lambda_j\lambda_k}Cov(\epsilon_{ij}, \epsilon_{ik}) = 0$. The first term can therefore be expressed as

$$
\begin{aligned}
&\mathbb{E}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}(\eta_i^1 + \epsilon'_{ij})]^2 \\
&= \sum_{j=1}^{k} (\frac{\omega_j \lambda_j}{\hat{\lambda}_j})^2 [(\eta_i^1)^2 + \frac{1}{\lambda_j^2}\sigma^2(\epsilon_{ij})] + \sum_{j \neq l} (\frac{\omega_j \lambda_j}{\hat{\lambda}_j}\frac{\omega_l \lambda_l}{\hat{\lambda}_l})(\eta_i^1 \eta_i^1 + \frac{1}{\lambda_j \lambda_l} Cov(\epsilon_{ij}, \epsilon_{il})) \\
&= [\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\eta_i^1]^2 + \mathbb{E}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\epsilon'_{ij}] \\
&= [\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\eta_i^1]^2 + \sum_{j=1}^{k} \frac{\omega_j}{\hat{\lambda}_j}\sigma^2(\epsilon_{ij}) \quad \textbf{if } Cov(\epsilon_{ij}, \epsilon_{ik}) = 0
\end{aligned}
$$

This result implies that we can separate η and measurement error when the measurement errors for each measure are independent of one another.

Similarly, we observe that in the second part, η and ϵ can also be separated:

$$
\begin{aligned}
\mathbb{E}[S_1^2] &= \frac{1}{n-1}\sum_{i=1}^{n}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\eta_i^1 - \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j})\eta_i^1]^2 \\
&\quad + \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\epsilon'_{ij} - \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\epsilon'_{ij})]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\eta_i^1 - \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j})\eta_i^1]^2 \\
&\quad + Var(\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\epsilon'_{ij}) \quad \text{b/c expectation of sample variance is unbiased} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j}\eta_i^1 - \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{k} \frac{\omega_j \lambda_j}{\hat{\lambda}_j})\eta_i^1]^2 \\
&\quad + \sum_{j=1}^{k} \frac{\omega_j}{\hat{\lambda}_j}\sigma^2(\epsilon_{ij})
\end{aligned}
$$

The first line is the variance of weighted average of potential outcome η without measurement error, and the second line is the variance weighted average of measurement error.

It is important to note that the sample variance analogue $\hat{S}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} Z_i(\tilde{Y}_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i)^2$ and $\hat{S}_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(1-Z_i)(\tilde{Y}_i - \frac{1}{n_1}\sum_{i=1}^{n}(1-Z_i)\tilde{Y}_i)^2$ is unbiased for σ_0^1 and σ_0^0 [22]:

[22]The third line can be verified by expanding the quadratic terms and checking one by one. For the first term, it is

$$\begin{aligned}\mathbb{E}[\hat{S}_1^2] &= \mathbb{E}_\epsilon[\mathbb{E}_Z(\hat{S}_1^2|\epsilon)] \\ &= \mathbb{E}_\epsilon[\mathbb{E}_Z[\frac{1}{n-1}\sum_{i=1}^{n} Z_i(\tilde{Y}_i - \frac{1}{n_1}\sum_{i=1}^{n} Z_i\tilde{Y}_i)^2]] \\ &= \mathbb{E}_\epsilon[S_1^2]\end{aligned}$$

The variance in the super-population framework follows similar results for a completely randomization experiment. See Imbens and Rubin (2015). □

B.4 Proof of Proposition 4

Proof. Define the OLS estimator

$$\begin{bmatrix}\hat{\alpha} \\ \hat{\beta}\end{bmatrix} = (X'X)^{-1}X'[\tilde{Y}]$$

Note that

$$(X'X) = \begin{bmatrix} Jn & Jn_1 \\ Jn_1 & Jn_1 \end{bmatrix} \tag{8}$$

Let T denote the index set of treated individuals, and C denote the index set of control individuals. Then, algebra shows that

$$(X'X)^{-1}X'[\tilde{Y}] = \frac{1}{J^2 n_1(n-n_1)}\begin{bmatrix} Jn_1 & -Jn_1 \\ -Jn_1 & Jn \end{bmatrix}\begin{bmatrix} \sum_{i=1}^{n}\sum_{j=1}^{J}\tilde{Y}_{ij} \\ \sum_{i\in T}\sum_{j=1}^{J}\tilde{Y}_{ij} \end{bmatrix} \tag{9}$$

clear that, because $Z_i(1-Z_i)=0$,

$$\begin{aligned}\mathbb{E}[Z_i\tilde{Y}_i^2] &= \mathbb{E}[Z_i\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(Z_i\eta_i^1 + (1-Z_i)\eta_i^0 + \epsilon'_{ij})]^2 \\ &= \mathbb{E}[\sum_{j=1}^{k}\frac{\omega_j\lambda_j}{\hat{\lambda}_j}(Z_i\eta_i^1 + Z_i\epsilon'_{ij})]^2 \\ &= \mathbb{E}[Z_i(\tilde{Y}_i^1)^2]^2\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\beta} &= \frac{1}{kn_1}\sum_{i\in T}\sum_{j=1}^{J}\tilde{Y}_{ij} - \frac{1}{kn_0}\sum_{i\in C}\sum_{j=1}^{J}\tilde{Y}_{ij} \\ &= \frac{1}{n_1}\sum_{i\in T}(\frac{1}{k}\sum_{j=1}^{J}\tilde{Y}_{ij}) - \frac{1}{n_0}\sum_{i\in C}(\frac{1}{k}\sum_{j=1}^{J}\tilde{Y}_{ij}) \\ &= \frac{1}{n_1}Z_i\sum_{i=1}^{n}(\frac{1}{k}\sum_{j=1}^{J}\tilde{Y}_{ij}) - \frac{1}{n_0}(1-Z_i)\sum_{i=1}^{n}(\frac{1}{k}\sum_{j=1}^{J}\tilde{Y}_{ij})\end{aligned}$$

□

B.5 Proof of Proposition 5

Proof. Define Centered variables $X_{ij} = Y_{ij} - L_j(\eta_i, \alpha_j)$, so $\mathbb{E}X_{ij} = 0$. Let $S_J = \sum_{j=1}^{J} X_{ij}$. Therefore, the condition that $\sum_{J=1}^{\infty} \frac{\sum_{j=1}^{J}\sum_{k=1}^{J} Cov(Y_{ij}, Y_{ijk})}{J^2} < \infty$ is equivalent to $\sum_{J=1}^{\infty} \frac{Var(S_J)}{J^2} < \infty$

By Chebyshev's inequality, for any $\epsilon > 0$, $\mathbb{P}[|S_J| \geq \epsilon J] \leq \frac{Var(S_J)}{\epsilon^2 J^2}$. Hence, $\sum_{J=1}^{\infty}\mathbb{P}[|S_J| \geq \epsilon J] \leq \sum_{J=1}^{\infty}\frac{Var(S_J)}{\epsilon^2 J^2}$. By assumption of $\sum_{J=1}^{\infty}\frac{Var(S_J)}{J^2} < \infty$, we confirm $\sum_{J=1}^{\infty}\mathbb{P}[|S_J| \geq \epsilon J] < \infty$. Next, by Borel-Cantelli lemma, we conclude $\lim_{J\to\infty}\frac{S_J}{J} = 0$ a.s., which implies $\overline{Y}_i \to_{a.s.} \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J} L_j(\eta_i, \alpha_j)$. If the limit is well-define, it is still a linear function of η_i. □

C The Consequences of Sampling Variability When Estimating Scaling Factors

Scaling factors λ_j are sometimes known based on previous research, but in many cases they are estimated using the experimental data at hand. In that case, two-step estimators are potentially problematic because sampling uncertainty associated with the first step (estimating the scaling parameters) must be taken into account when estimating the standard errors in the second step (estimating the ATE on the latent outcome, measured by a weighted index).

How serious is the failure to acknowledge the uncertainty of $\hat{\lambda}_j$ from the first step? In this section, we show that the variance of $\hat{\lambda}_j$ tends to be small. For example, we can estimate $\hat{\lambda}_j$ by $\hat{\lambda}_j = \frac{Cov(Z,Y_j)}{Cov(Z,Y_1)}$. The scaling parameter can also be estimated by the ratio of two OLS estimators $\frac{\beta_{Y_j\sim Z}}{\beta_{Y_1\sim Z}}$, where the numerator is the OLS coefficient for the regressor Z in the regression of Y_j on Z; the denominator is similar, except now the regression is Y_1 on Z. By a Taylor series expansion, $\frac{\beta_{Y_j\sim Z}}{\beta_{Y_1\sim Z}} \approx \lambda_j + (\beta_{Y_j\sim Z} - \lambda_j) - \lambda_j(\beta_{Y_1\sim Z})$

When the sample size is large, the variance can be approximated by

$$
\begin{aligned}
Var(\hat{\lambda}_j) &\approx Var(\beta_{Y_j \sim Z}) + \lambda_j^2 Var(\beta_{Y_1 \sim Z}) \\
&\approx \frac{\sigma^2(\epsilon_j)}{\sum_{i=1}^n (Z_i - \overline{Z})^2} + \frac{\lambda_j^2 \sigma^2(\epsilon_1)}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \\
&\approx \frac{4}{n}[\sigma^2(\epsilon_j) + \lambda_j^2 \sigma^2(\epsilon_1)]
\end{aligned}
$$

This approximation tells us that the variance of the $\hat{\lambda}_j$ is a weighted sum of measure error variances but declines with n and therefore will be small in practice when n is large.

D Further Thoughts on Identifying Scaling Factors

It is instructive to examine the first stage of the IV estimators of the scaling parameters. Suppose Z_i is the instrument, the first stage is

$$
\begin{aligned}
Y_{i1} &= Z_i \eta_i^1 + (1 - Z_i)\eta_i^0 + \epsilon_{i1} \\
&= (\eta_i^0 + Z_i \tau_i) + \epsilon_{i1} \\
&= \mathbb{E}[\eta_i^0] + \mathbb{E}[\tau_i] Z_i + (\eta_i^0 - \mathbb{E}[\eta_i^0] + (\tau_i - \mathbb{E}[\tau_i] Z_i) + \epsilon_{i1}) \\
Y_{i1} &:= \mathbb{E}[\eta_i^0] + \mathbb{E}[\tau_i] Z_i + \xi_i
\end{aligned}
$$

Under the assumption of random assignment in Assumption 1, $\mathbb{E}[\xi_i | Z_i] = 0$, which in turn implies the unbiasedness of the least squares estimator in the first stage.

In the case where Y_{i3} serves as the instrumental variable, from the perspective of two-stage least squares, the first stage is

$$
Y_{i1} = \frac{1}{\lambda_3} Y_{i3} + (\epsilon_{i1} - \frac{1}{\lambda_3} \epsilon_{i3})
$$

Although in the structural linear model Y_{i3} is correlated with ϵ_{i3}, we can always form a linear projection with $\tilde{\lambda}_3$ and $\tilde{\epsilon}_{i3}$ such that

$$
Y_{i1} = \tilde{\lambda}_3 Y_{i3} + \tilde{\epsilon}_{i3}
$$

and $\mathbb{E}[\tilde{\epsilon}_{i3} Y_{i3}] = 0$.[23]

[23] For further discussion of linear projection, see Hansen (2022).

E GMM

Consider the example with three measures. We seek to estimate λ_2, λ_3, and τ. Because λ_2, λ_3 are over-identified through multiple IVs, as we shown in Proposition 1, the moment conditions are:

$$\mathbb{E}\begin{bmatrix} Y_{i2} - \alpha_0 - \lambda_2 Y_{i1} \\ (Y_{i2} - \alpha_0 - \lambda_2 Y_{i1}) Z_i \\ (Y_{i2} - \alpha_0 - \lambda_2 Y_{i1}) Y_{i3} \\ Y_{i3} - \alpha_1 - \lambda_3 Y_{i1} \\ (Y_{i3} - \alpha_1 - \lambda_3 Y_{i1}) Z_i \\ (Y_{i3} - \alpha_1 - \lambda_3 Y_{i1}) Y_{i2} \\ (\frac{Z_i}{p} - \frac{1-Z_i}{1-p}) Y_i - \tau \end{bmatrix} = 0$$

The first three equations are over-identified IV equations that estimate λ_2. The last equation is the weighted-average estimator for the average treatment effect, where p is the propensity score. Then, the variance of τ is estimated through traditional sandwich estimators that incorporate the uncertainty in the estimate of λ_j. The last two rows can be replaced by $[(Y_{i1} + \frac{Y_{i2}}{\lambda_2} + \frac{Y_{i3}}{\lambda_3})/3 - \mu_1] Z_i$ and $[(Y_{i1} + \frac{Y_{i2}}{\lambda_2} + \frac{Y_{i3}}{\lambda_3})/3 - \mu_0](1 - Z_i)$, so that the difference-in-means estimator is expressed in two parts rather than one: $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$ and the variance is $Var(\mu_1) + Var(\mu_0) - 2Cov(\mu_1, \mu_0)$.

For an optimally weighted scaled index, because we need to deal with covariance, we centered the outcome measures first. The following Y_{ij} is assumed to be demeaned first.

$$\mathbb{E}\begin{bmatrix} (Y_{i2} - \lambda_2 Y_{i1}) Z_i \\ (Y_{i2} - \lambda_2 Y_{i1}) Y_{i3} \\ (Y_{i3} - \lambda_3 Y_{i1}) Z_i \\ (Y_{i3} - \lambda_3 Y_{i1}) Y_{i2} \\ Y_{1i} Y_{2i} - \lambda_2 \psi \\ Y_{1i}^2 - \psi - \sigma_1^2 \\ Y_{2i}^2 - \lambda_2^2 \psi - \sigma_2^2 \\ Y_{3i}^2 - \lambda_3^2 \psi - \sigma_3^2 \\ (\frac{Z_i}{p} - \frac{1-Z_i}{1-p}) Y_i^{opt} - \tau \end{bmatrix} = 0$$

In the above moment conditions, ψ is the variance of η, σ_j^2 is the variance of $\epsilon_{\cdot j}$, $Y_i^{opt} = \sum_{j=1}^{J} \omega_j Y_{ij}$, and $\omega_j = \frac{\lambda_j^2/\sigma_j^2}{\sum_{j=1}^{k} \lambda_j^2/\sigma_j^2}$.

A preliminary R package can be found at Github.

F Pseudo-algorithm for Difference-in-means based on a Weighted Scaled Index

1. **Data:** Z_i, Y_{ij} $(i = 1, 2, ..n; j = 1, 2, ..., K)$, $n_1 = \sum_{i=1}^{n} Z_i$, $n_0 = n - n_1$
2. **Estimate** λ_j: set $\lambda_1 = 1$, for $\lambda_j \neq \lambda_1$, use either of the following estimators:
   (1) SEM: $\eta =\sim 1 * Y_1 + Y_2 + ... + Y_n$; or
   (2) GMM: using Z_i or/and $Y_{ik} \forall k \neq 1$ and $k \neq j$ as IVs
3. **Scale outcomes:** for each i, j, $\tilde{Y}_{ij} \leftarrow \frac{Y_{ij}}{\hat{\lambda}_j}$
4. **Weight outcomes:** $\tilde{Y}_i \leftarrow \sum_{j=1}^{K} \omega_j \tilde{Y}_{ij}$, where $\sum_{i=1}^{K} \omega_j = 1$. Common options:
   (1) Uniform weights: $\omega_j \leftarrow \frac{1}{K}$
   (2) Optimal weights: $\omega_j \leftarrow \frac{\hat{\lambda}_j^2 / \hat{\sigma}^2(\epsilon_j)}{\sum_{j=1}^{k} \hat{\lambda}_j^2 / \hat{\sigma}^2(\epsilon_j)}$
5. **Output:** $\hat{\tau} \leftarrow \frac{1}{n_1} \sum_{i=1}^{n} Z_i \tilde{Y}_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) \tilde{Y}_i$

G Further Evidence that Additive Indices Help Linearize the Relationship between the Observed and Latent Outcomes

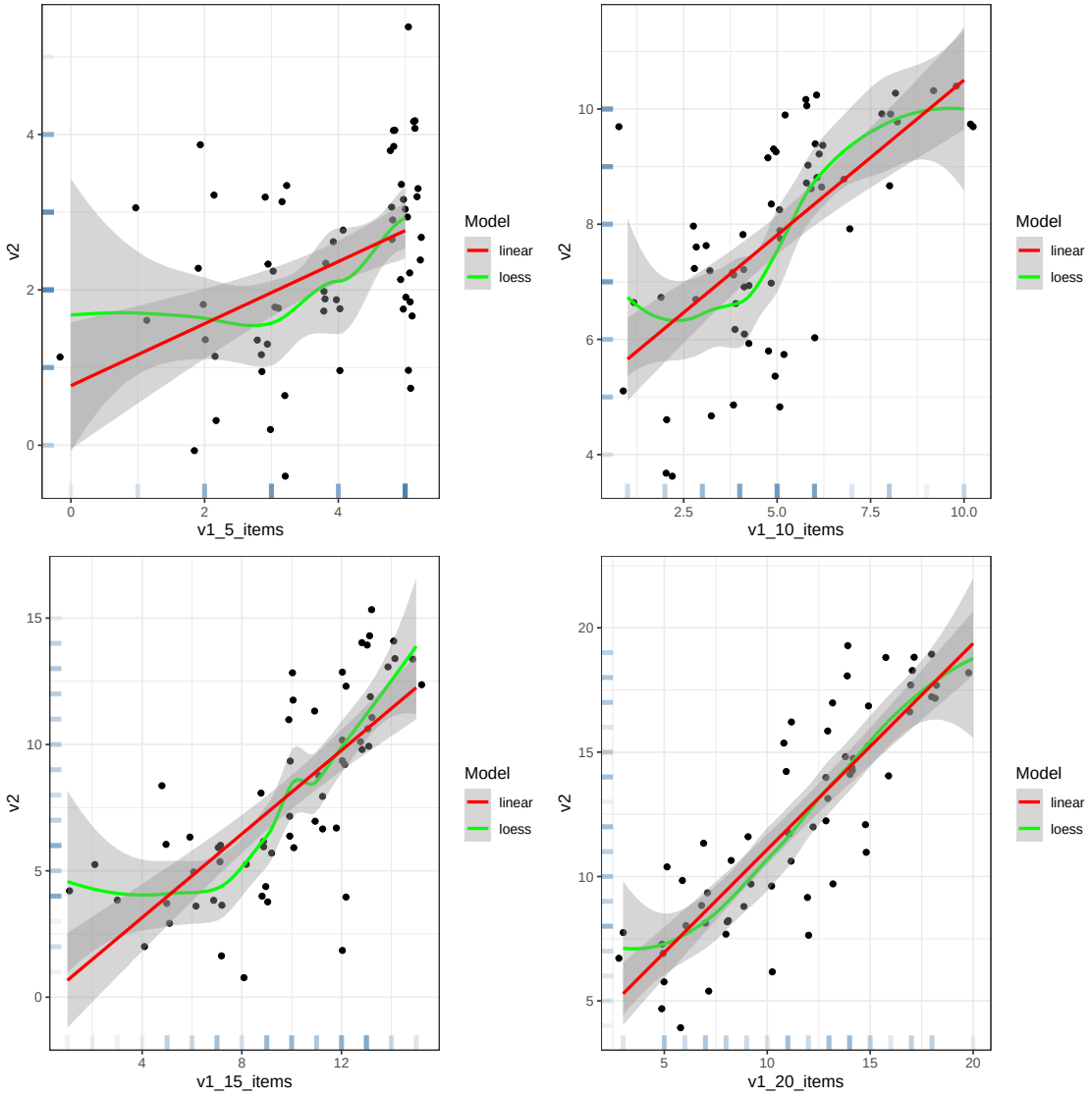

Figure A.2: **Linearity between measurements by adding binary items together**. In each figure, we use the same data from figure 2. We create two variables v_1 and v_2 by summing up 5, 10, 15, and 20 binary responses from IRT models. Data points are jittered. The rug plots on both axes denote the distribution of data.

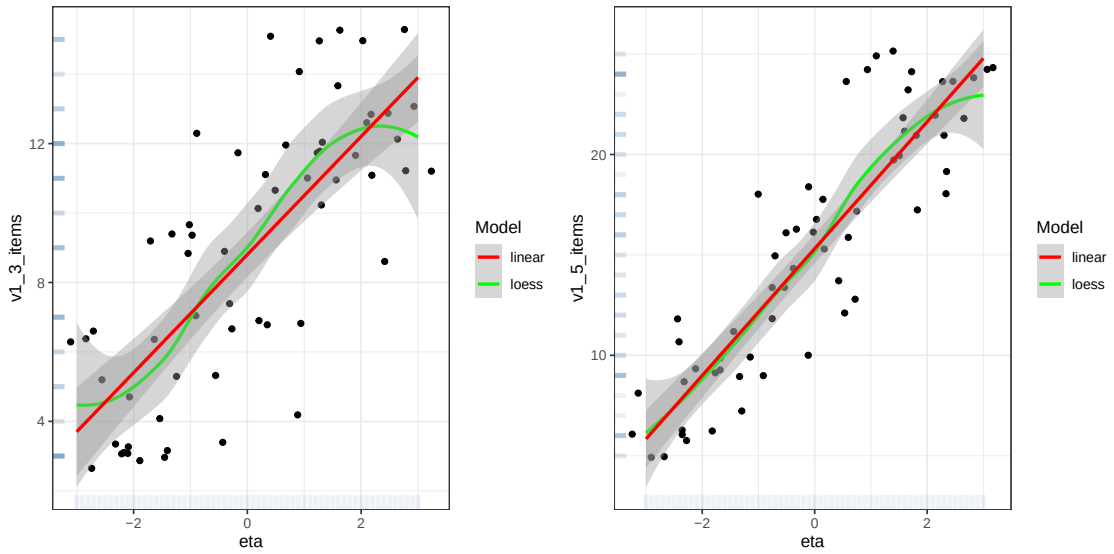

Figure A.3: **Linearity between η and measurements by adding 5-point ordinal items together**. We create an additive index v_1 by summing up 3 and 5 ordinal responses (taking values 1, 2, 3, 4, and 5) generated from IRT models. Data points have been jittered slightly for clarity. The rug plots on both axes denote the distribution of data. The vertical axis shows the additive index created by adding all ordinal variables in the simulation. The horizontal axis is the true latent variable (η) used in the IRT model. The figures show that even a relatively small number of 5-point scales produce an index that bears an approximately linear relationship to the latent variable.

H Comparison to current practice

In social science, it is quite common for a study's key outcome variable to be an abstract concept that is not directly observed. To better measure this latent variable, researchers often seek to measure multiple observable manifestations of this latent outcome. However, researchers are aware that critics may express concern about false discovery due to multiple comparisons. To sidestep this concern, researchers often attempt to reduce the dimensionality of their outcomes.

Common dimensionality reduction techniques include summary index creation, principal components analysis, or other techniques that in some way standardize the latent and/or observed outcomes. A key feature of those methods is that they extract or construct a low-dimensional, typically single-dimensional, variable. Due to standardization (e.g., recoding each outcome measure to have a variance of 1), these methods generate results that are sample-specific in the sense that the scaling used to define the outcome variable depends on the dispersion of scores among the subjects at hand. Even if the true causal relationship between treatment and outcome were identical in two different (large) samples, estimated effects might differ substantially if the dispersion of unobservables differs. As noted in the main text, our method solves the above problem by setting the metric of the latent variable to be the same as one of the observed variables. This "unstandardized approach" allows us to interpret the latent variable with the same metric as that observed outcome variable (see Bollen (1989, pp. 239-240) and Loehlin (1998, pp. 28-29)). The results are comparable across samples even when the variances of the outcome variables differ.

Another approach, as noted in the main text, is to simply consider observed outcomes as distinct from one another, even if they are in the same substantive domain. There is no latent variable in this case; just multiple outcomes. In his widely-cited article Anderson (2008) seeks to find a weight vector $w \in R^J$ so that the variance of the weighted outcome $wY \in R^{n\times 1}$ is minimized, subject to the constraint that $1'w = 1$. This is an optimization problem, where the objective function is $\min_w w'\Sigma w$. By using the Lagrange multiplier method, one obtains the weight $w = (1'\Sigma^{-1}1)^{-1}(1'\Sigma)$.

How well does Anderson's inverse-covariance weighting (ICW) method perform in terms of statistical power? (We focus on power because we know that Anderson's standardized index will not yield unbiased or consistent estimates of the ATE in unstandardized units).

In our simulation, we set $\eta_{i0} = 0$, $\eta_{i1} = N(\theta, 1)$, where $\theta \in \{0, 0.05, 0.15, 0.25, 0.35, 0.45\}$. For three outcome measures, we let $\lambda_j = 1$ and measurement errors $e_1 = N(0, 0.5)$,$e_1 = N(0, 0.1)$, and $e_1 = N(0, 2)$. Figure A.4 illustrates the power of equal/optimal weighting estimator, the SEM estimator, and Anderson's ICW estimator. The optimal weighting estimator and SEM estimator have the greatest power (they are almost indistinguishable in the figure), dominating ICW. The reason is that for SEM and optimal weighting estimators, the largest weights are assigned to mea-

sures with smallest measurement error, which increases the overall precision. However, the ICW procedure down-weights outcomes that covary with other outcomes so that the highest weight is not necessarily assigned to the most precise measure.

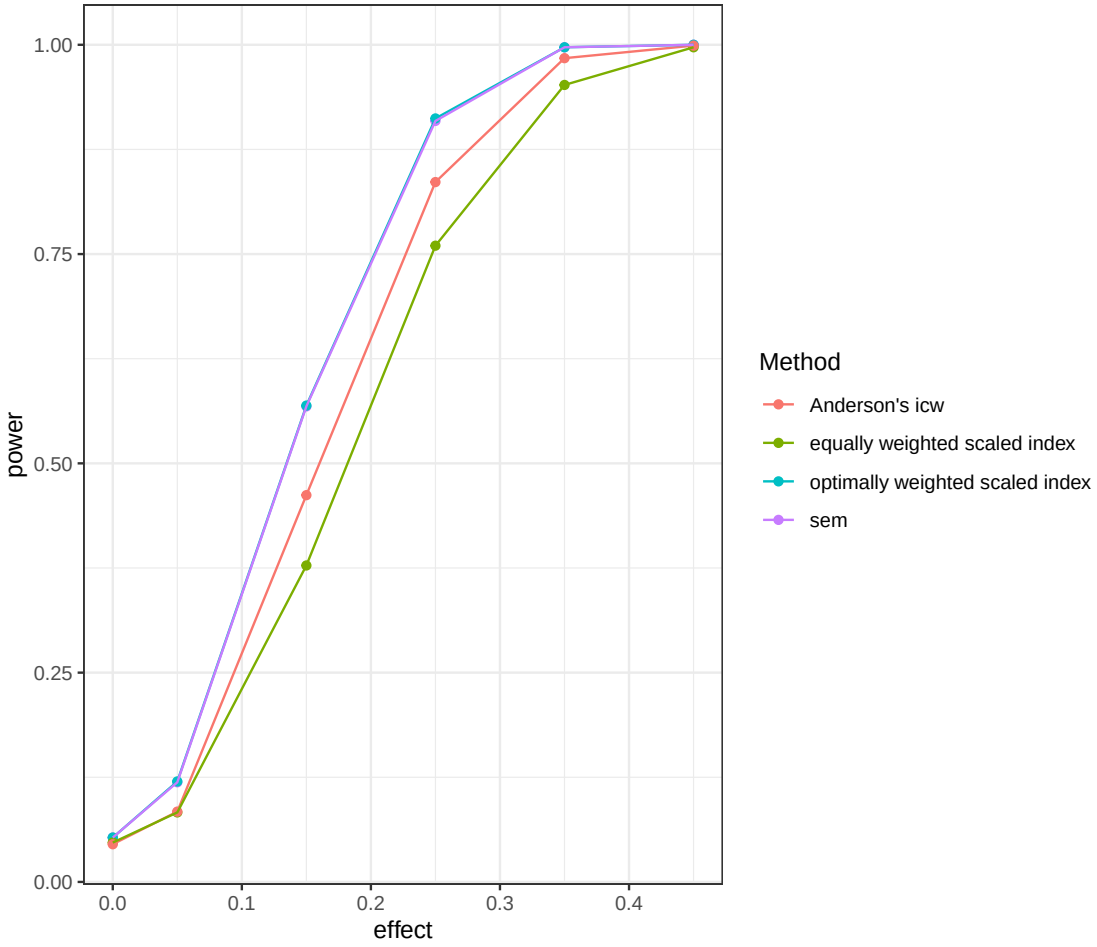

Figure A.4: Power analysis: Optimal Weighing, SEM, and ICW

The next simulation compares ICW and SUR (based on the F-test). We generate data from the DGP in Figure A.5. We examine two situations, low correlation (0.2) and high correlation (0.8) among the disturbances. ICW is somewhat more powerful than SUR in both scenarios.

I Comparison to Hierarchical Item Response Theory

Stoetzer et al. (2022) propose a hierarchical item response theory (hIRT) model to estimate the latent treatment effect. Their model has two components. First, the latent variable η_i is assumed to be a linear function of the treatment: $\eta_i = \gamma_0 + \gamma_1 Z_i + \epsilon_i$, where ϵ_i is normally distributed with a constant variance σ^2. In this equation, γ_1 is the average treatment effect of interest. Next,

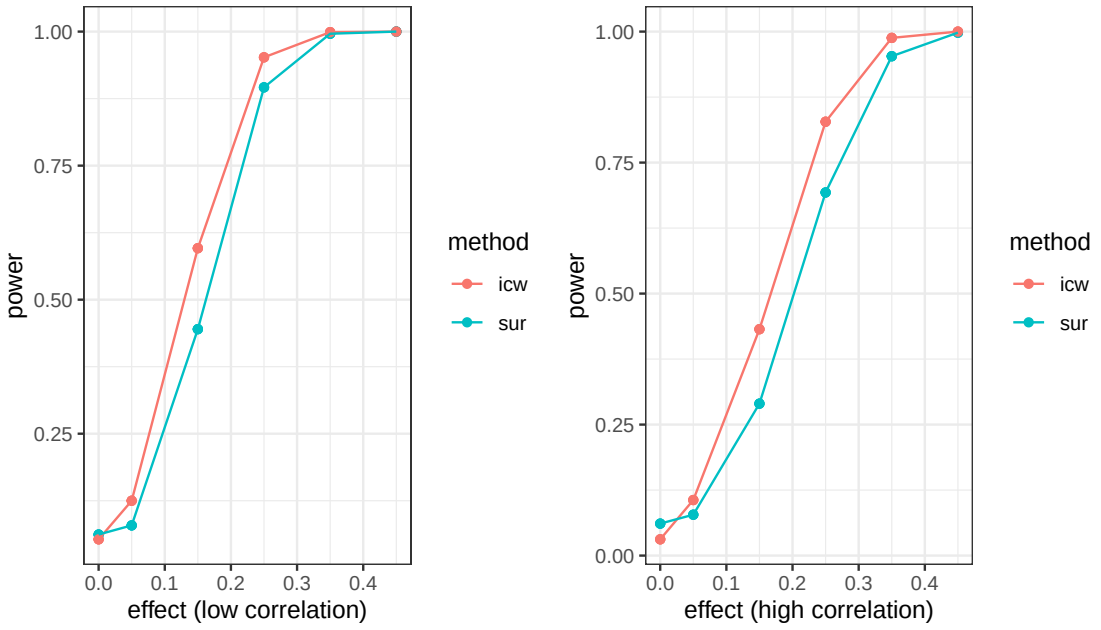

Figure A.5: Power analysis: SUR and ICW

the observed outcomes Y_{ij} are generated by a characteristic function of the item: $\mathbb{P}[Y_{ij} = h|\eta_i] = P_{ijh}(\eta_i; \alpha_{ijh}, \beta_{ij})$, where α_{ijh} and β_{ij} are the item difficulty and item discrimination parameters for item j and answering h.

As Stoetzer et al. (2022) note (e.g. page 28), this model makes a set of parametric assumptions, including a constant treatment effect and a particular item characteristic function. Our approach relaxes some of these assumptions. We allow Z_i to have heterogeneous treatment effects. We also remain agnostic about the non-linear transformation function g.

To estimate γ_1 under the hIRT, several other identification restrictions must be imposed. For example, $\lambda_0 = 0$ and $\sigma^2 = 1$. Then, γ_1 can be estimated by the EM algorithm. These technical assumptions have the same normalization purpose as setting $\lambda_1 = 1$. However, setting $\lambda_1 = 1$ allows us to give a natural interpretation to the latent variable.

An immediate shortcoming of the parametric IRT approach is that the estimator is likely to be inconsistent when the parametric function form is misspecified or there are heterogeneous effects. To illustrate the problem, we conduct two simulations. For constant effect, we generate potential outcomes $\eta_{i0} = N(0, 1)$ and $\eta_{i1} = 2 + \eta_{i0}$ so that treatment effect is exactly 2 for every individual. For heterogeneous effects, we simply add some mean zero noise: $\eta'_{i1} = 2+\eta_{i0}+N(0, 1)$. Note that the average treatment effect is still maintained at 2. Next, we generate five measures (each with 13 levels) using an IRT model. Figure A.6 illustrates the mean estimate and its 95% confidence interval. When the effect of treatment is constant for every individual, hIRT recovers the true effect of 2. However, after introducing even mild heterogeneity in treatment effects, hIRT exhibits

substantial bias.

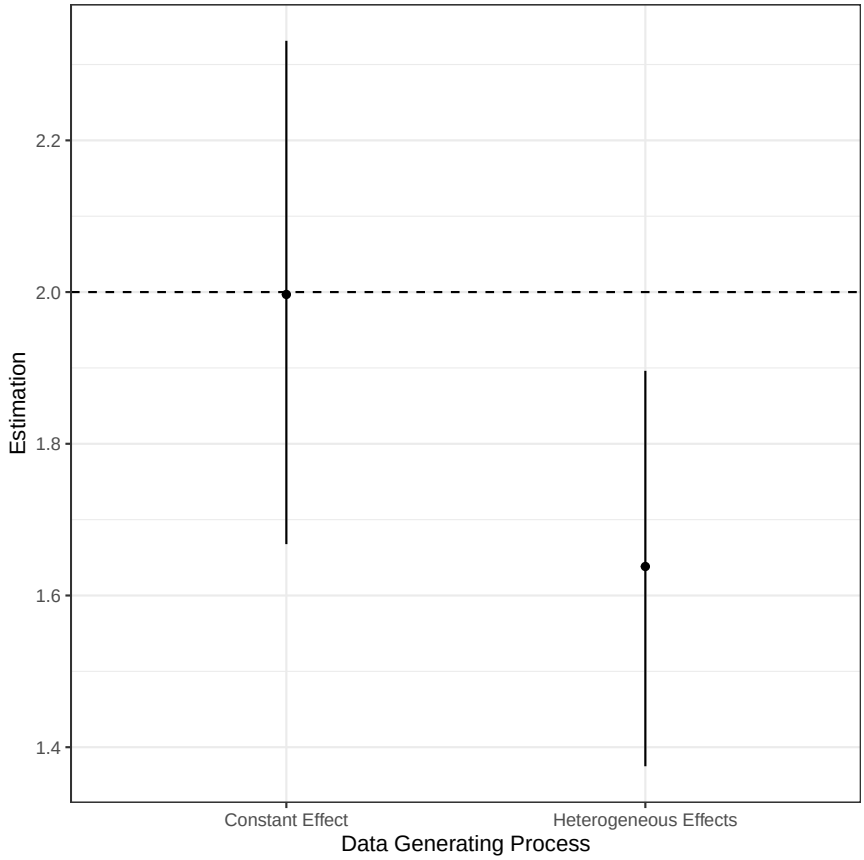

Figure A.6: The Distribution of hIRT Estimates under Constant Treatment Effects and Heterogeneous Treatment Effects.

J More Information on the Empirical Application

J.1 Data

Here we list the outcome measurements and covariates we used in the application section.

Anti-Immigrant Prejudice Index. The first set of questions are five point scales where respondents were asked: "Do you agree or disagree with the below statements about undocumented or illegal immigrants?" Response options were: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree:

- 'living': "I would have no problem living in areas where undocumented immigrants live."
- 'fit': "Too many undocumented immigrants just don't want to fit into American society."

- ‘burden’: ”Undocumented immigrants are too much of a burden on our communities.”
- ‘crime’: ”Undocumented immigrants have already broken the law coming here illegally, so they are more likely to commit other crimes.”
- ‘values’: ”Undocumented immigrants hold the same values as me and my family.”

Anti-Immigrant Policy Index. Respondents were first asked: ”Politicians are considering a number of policies about immigration. We want to know what you think. Do you agree or disagree with the statements below?” Response options were: Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree:

- ‘daca’: ”The federal government should grant legal status to people who were brought to the US illegally as children and who have graduated from a U.S. high school.”
- ‘citizenship’: ”The federal government should allow undocumented immigrants currently in the U.S. to become citizens after they have lived, worked, and paid taxes for at least 5 years.”
- ‘compassion’: ”Undocumented immigrants deserve compassion and should not live in daily fear of deportation.”

Baseline covariates. They include the pre-treatment variables ‘daca’,‘citizenship’,‘living’ and ‘fit’. We add them together to construct a single index.

J.2 Supplementary results

Table A.1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Full Treatment	7870	0.33	0.47	0	0	1	1
Abbreviated Treatment	7870	0.33	0.47	0	0	1	1
daca	1578	0.6	1.4	-2	-1	2	2
citizenship	1578	-0.14	1.5	-2	-2	1	2
compassion	1578	-0.34	1.5	-2	-2	1	2
living	1578	0.51	1.3	-2	0	2	2
values	1578	0.58	1.2	-2	0	2	2
fit	1578	-0.4	1.4	-2	-2	1	2
burden	1578	-0.27	1.4	-2	-2	1	2
crime	1578	-0.87	1.2	-2	-2	0	2

Table A.2: Covariance Matrix

	Treatment (full)	Treatment (mod)	Attitudes	Policy Views	Covariates
Treatment (full) (Ave: 0.33)	0.226	-0.109	0.061	0.104	-0.035
Treatment (mod) (Ave: 0.33)	-0.109	0.216	0.008	0.010	0.052
Attitudes (Ave: 2.06)	0.061	0.008	12.328	15.187	12.775
Policy Views (Ave: 2.63)	0.104	0.010	15.187	30.222	19.782
Covariates (Ave: 1.76)	-0.035	0.052	12.775	19.782	19.318

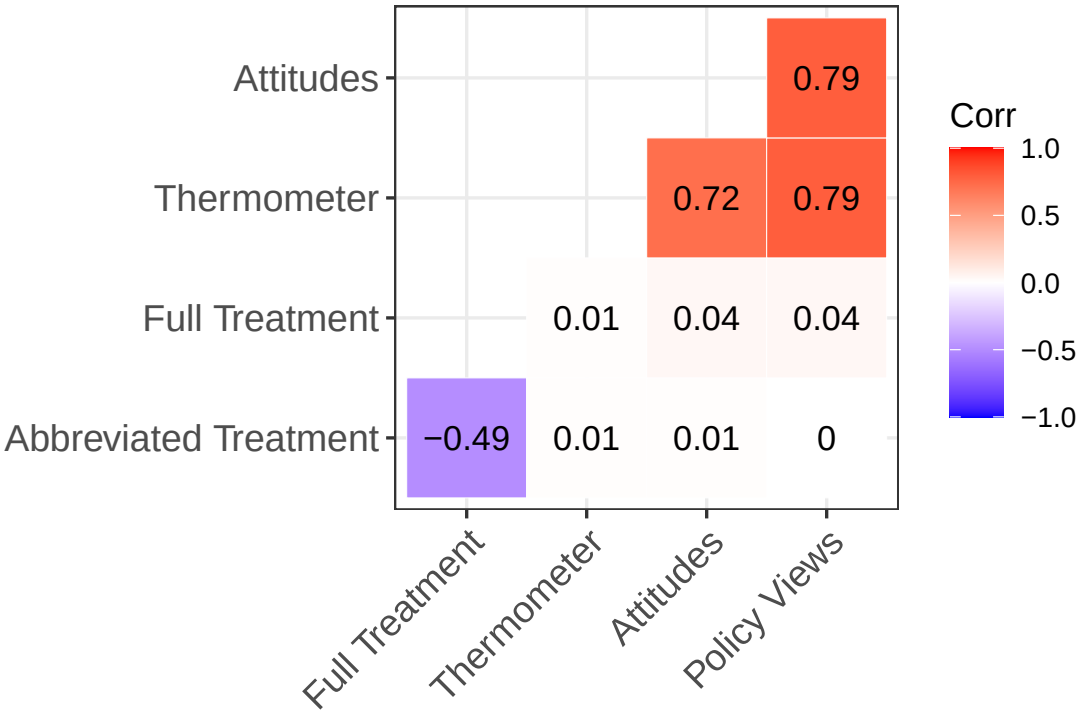

Figure A.7: Correlation Matrix

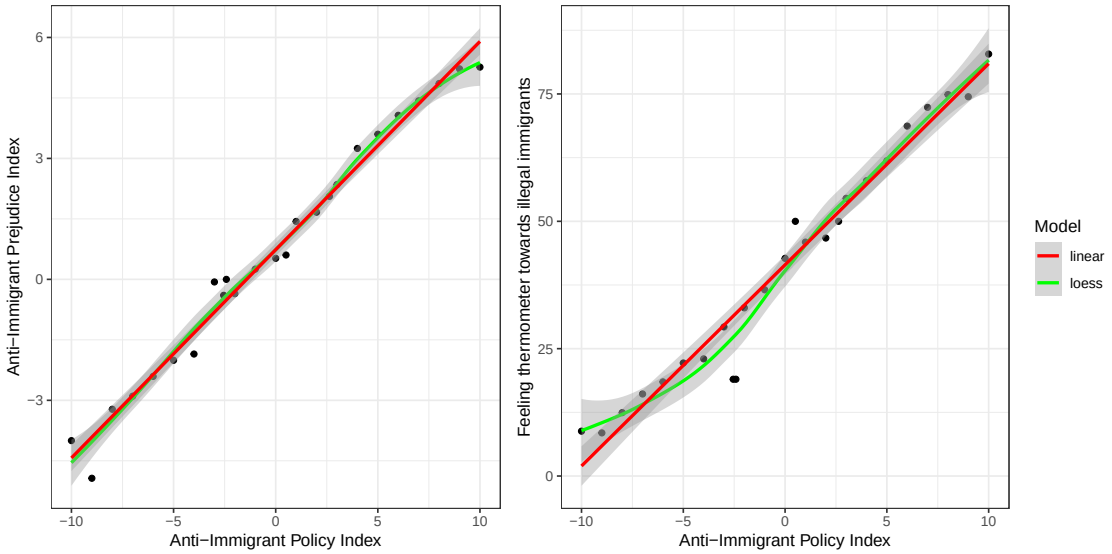

Figure A.8: Linearity check

Table A.3: Variance Estimation

	Model without Baseline covariates			
Variables	Var est	se	z	pvalue
.Attitudes	3.14	3.24	0.97	0.33
.Policy Views	5.11	8.86	0.58	0.56
Treat (full)	0.23	0.01	28.09	0.00
Treat (mod)	0.22	0.01	28.09	0.00
.eta	9.16	3.26	2.81	0.00
	Model with Baseline covariates			
.Attitudes	2.52	0.13	18.97	0.00
.Policy Views	6.70	0.33	20.00	0.00
Treat (full)	0.23	0.01	28.09	0.00
Treat (mod)	0.22	0.01	28.09	0.00
Cov	19.31	0.69	28.09	0.00
.eta	1.32	0.11	12.17	0.00

Note: This table shows the variance estimation of observed and latent variables in the application. The · in front of the parameter denotes residual variance for the variable.

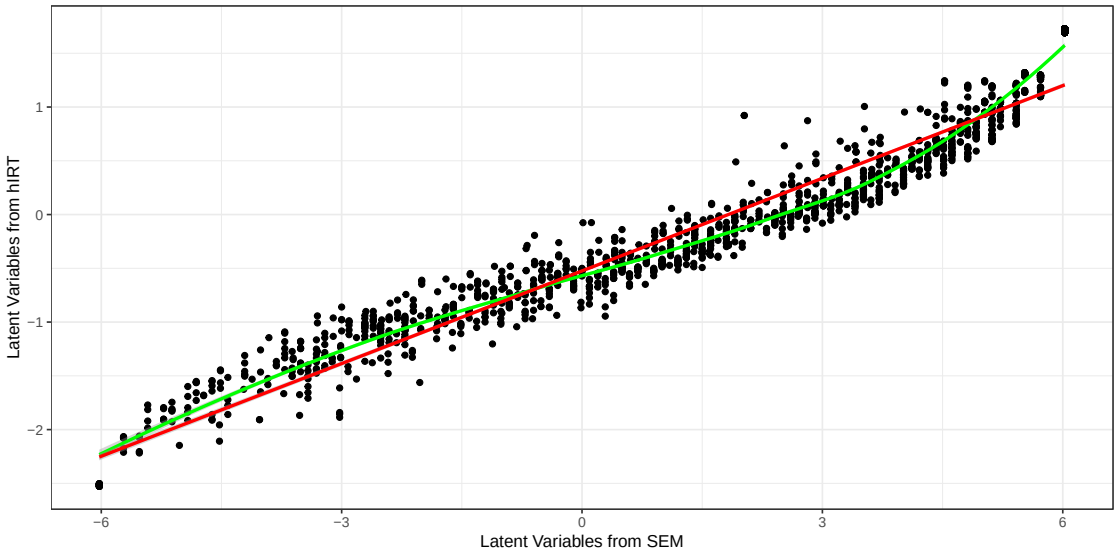

Figure A.9: Scatterplot of the Imputed Latent Variables implied by the SEM and hIRT Models, with Linear and LOESS Fitted Lines

K Illustration of How Additional Items Reduce the Sampling Variability of the Estimated ATE

In the simulation, we generate $n = 500$ potential outcomes $\eta^1 \sim N(1,1)$ and $\eta^0 \sim N(2,1)$, and 6 outcome measures $Y_{ij} = Z\eta_i^1 + (1-Z_i)\eta_i^0 + \epsilon_i$. We consider two cases, high reliability and low reliability, brought about by changing the variance of the measurement error. The average of 1000 simulation results is shown in Figure A.10 and Table A.4. Generally, more measures decrease the estimation variance. The benefits of additional measures depend on the measures' reliability. When measures are already quite reliable, the variance reduction is small.

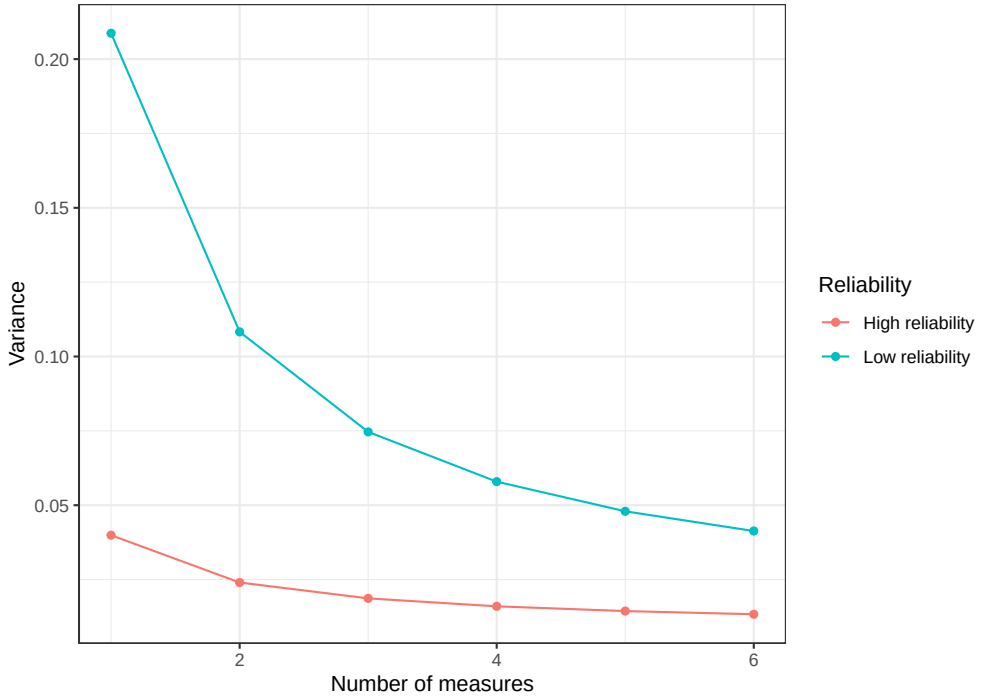

Figure A.10: **Simulation: Variance reduction and Reliability.** The horizontal line represents the number of measures and the vertical line is the estimated variance of the optimal weighting estimator. The variance reduction is larger if the reliability is lower.

The first column denotes the number of measures and the third column is the estimated variance for the LTE. The fourth column calculates the $\Delta(J)$. Recall $\Delta(J)$ is the variance reduction by adding the $J(+1)^{th}$ measure. The last two columns compare the simulation ratio and the theoretical value $\frac{J}{J+2}$ we derived in the main text.

Table A.4: Simulation: Variance reduction and Reliability

Number	Reliability	Var	$\Delta(J)$	$\Delta(J+1)/\Delta(J)$	Theory
1	High	0.0399	-0.0159	0.3351	0.3333
2	High	0.0240	-0.0053	0.5011	0.5000
3	High	0.0187	-0.0027	0.5985	0.6000
4	High	0.0160	-0.0016	0.6645	0.6667
5	High	0.0144	-0.0011		
6	High	0.0133			
1	Low	0.2087	-0.1004	0.3350	0.3333
2	Low	0.1083	-0.0336	0.4973	0.5000
3	Low	0.0747	-0.0167	0.5957	0.6000
4	Low	0.0579	-0.0100	0.6657	0.6667
5	Low	0.0480	-0.0066		
6	Low	0.0413			

L Exclusion Restriction Violations Stemming from Invalid Measurement

Consider the DGP in the figure A.11. Y_{i2} is affected by treatment Z_i through other channels beyond the latent variable of interest η_i. Setting $\lambda_1 = 1$ to set the scale for η_i, we write

$$\begin{aligned} Y_{i1} &= (\eta_i^0 + Z_i\tau_i) + \epsilon_{i1} \\ Y_{i2} &= \lambda_2(\eta_i^0 + Z_i\tau_i) + \tilde{\lambda}_2(\tilde{\eta}_i^0 + Z_i\tilde{\tau}_i) + \epsilon_{i2} \\ &:= \lambda_2(\eta_i^0 + Z_i\tau_i) + \tilde{\varepsilon}_{i2} \end{aligned}$$

Because $\tilde{\varepsilon}_{i2}$ is correlated with Z_i, assumption 1C is violated. When $\tilde{\lambda}_2 \neq 0$, λ_2 is not identified. The root of the problem is that Z_i has a causal path to Y_{i2} other than through η_i. The covariance that Y_{i1} and Y_{i2} share reflects something other than their shared dependence on η_i; this covariance is affected also by the fact that Z_i affects both η_i and $\tilde{\eta}_i$.

This case illustrates a potential downside of misspecifying a latent variable model. Although a simple regression of Y_{i1} on Z_i would recover a meaningful causal effect, a latent variable model that ignores the backdoor path from Z_i to Y_{i2} through $\tilde{\eta}_i$ will produced biased estimates of the ATE of Z_i on η_i. Under what conditions would this kind of misspecification arise? One scenario occurs when Y_{i2} is an invalid measure of η_i insofar as it measures another latent trait as well, and this other trait is itself affected by the treatment Z_i. Another scenario occurs when the treatment triggers a response bias that affects some measures but not others. This kind of artifact might occur if the treatment were to affect both a trait (e.g., authoritarian attitudes) as well as a source of mismeasurement (e.g., acquiescence to agree/disagree questions).

If so, how can we mitigate the problem? One solution is to collect additional valid measures. We would expect the bias from the invalid measure to be diluted if there are more valid outcome variables. In the simulation (1000 times), $\eta^0 = 0$ and $\eta^1 \sim N(0,1)$. We create valid outcomes $Y_{ij} = Z_i\eta^1 + (1-Z_1)\eta^0 + N(0,1)$ and an invalid measure $Y_{ij} = Z_i\eta^1 + (1-Z_1)\eta^0 + N(0,1) + 0.05Z_i$. The result in the figure A.12 confirms that when the number of valid measures increases, the bias and the variance decrease. Another approach is to both collect more valid measurements and stipulate a more agnostic measurement model that allows some of the outcome measures to tap into more than one latent trait.

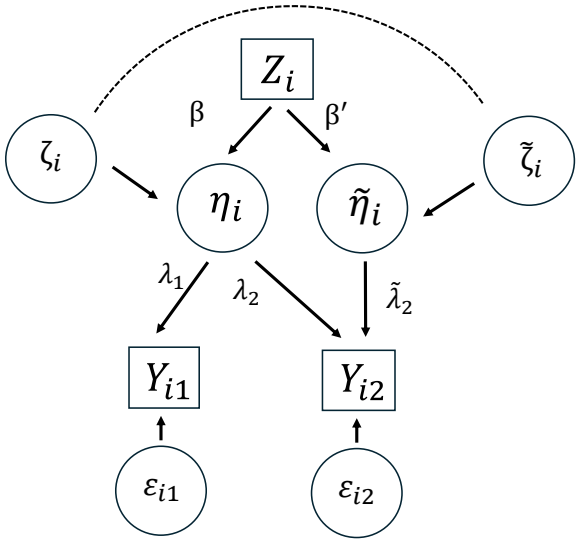

Figure A.11: Data generating process violating exclusion restriction.

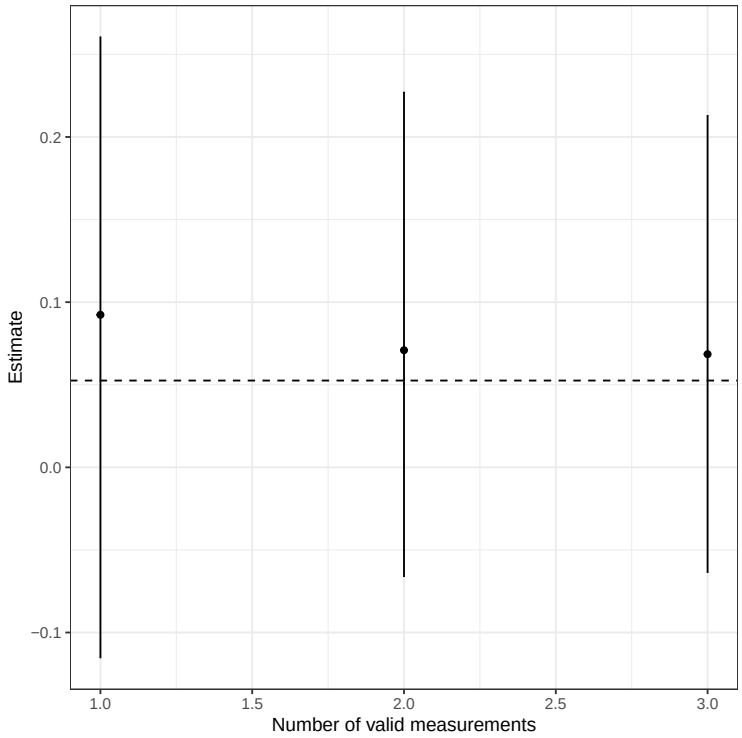

Figure A.12: Simulation: Adding Valid Measures Decreases Bias and Variance