NETWORK
LISER

Discussion Paper Series

IZA DP No. 18455

March 2026

How Does Tort Reform Affect Health Care Delivery

Charles Courtemanche
University of Kentucky, NBER and IZA@LISER

Joseph Garuccio
University of Kentucky

IZA@LISER NETWORK
ISSN: 2365-9793

IZA DP No. 18455 March 2026

How Does Tort Reform Affect Health Care Delivery[*]

Abstract

Health care costs in the U.S. have grown dramatically over the past several decades, with one possible cause being physicians providing unnecessary services out of fear of being sued for malpractice – a phenomenon known as “defensive medicine”. States responded by enacting different types of tort reforms. This paper reviews the literature on the effects of these tort reforms on outcomes related to malpractice risk, quantity and quality of health care services, overall utilization and expenditures, physician supply, and patient affordability. We use Google Scholar to identify papers that fall into this scope and use either associational or quasiexperimental quantitative methods. The preponderance of evidence points towards non-economic damage caps reducing malpractice risk, quantity of services (aside from diagnostics and obstetrics), and overall health care utilization and expenditures while increasing physician supply and not having detrimental effects on patient outcomes. In general, the effects of other types of tort reforms are less clear. The literature would benefit from further research utilizing recent methodological advances related to combining machine learning with causal inference and eliminating bias from heterogeneous treatment effects in staggered-treatment-time models.

JEL classification

I11, I18, K13, L25, L51

Keywords

tort reform, defensive medicine, physicians, hospitals, health care regulation

Corresponding author

Charles Courtemanche
courtemanche@uky.edu

[*] We acknowledge support for this research from the Kentucky Hospital Association under a grant entitled “An Economic Investigation of Regulations and Financial Challenges Facing Hospitals.” The views expressed in this paper are the authors’ and do not reflect those of the Kentucky Hospital Association. We thank Erin Osborne for excellent research assistance. Any errors are ours.

"Before beginning a Hunt, it is wise to ask someone what you are looking for before you begin looking for it."

– Winnie-the-Pooh

1 Introduction

Researchers in economics and related disciplines often wish to estimate empirical models containing variables corresponding to unobserved, multidimensional constructs. Such constructs are frequently incapable of being measured directly because they are inherently latent. As Dijkstra (2010, p. 25) observes, such *theoretical* concepts are "fundamental to the scientific enterprise in almost any field." Examples include college quality, corruption, economic freedom, financial literacy, human capital, health, institutional quality, permanent income, personality, political stability, quality of life, regulatory stringency, social capital, wealth, and many more (see, e.g., Ravallion 2012). Addressing this challenge has occupied researchers for more than a century, with different disciplines developing different norms.[1] The common thread across most solutions is to aggregate a set of observed variables—referred to as indicator or manifest variables—into a scalar composite index. After this dimensionality reduction, the index serves as a proxy for the latent construct in a regression model.[2]

The central econometric difficulty posed by latent constructs is measurement. Replacing an unobserved variable with a proxy—or a composite index—necessarily introduces measurement error, with implications for identification, estimation, and inference. More than four decades ago, Griliches (1985) emphasized that these issues are not peripheral but fundamental to empirical research. The most important message for practitioners, in our view, is that the *measurement model* is a structural component of research, as critical as the *economic model*. Griliches (1985, p. 198) argued:

> Theorists and model builders often proceed ... on the assumption that ideal data will be available and define variables that are unlikely to be observable, at least not in their pure form. Nor do they specify in adequate detail the connection between the actual numbers and their theoretical counterparts... [A]ny serious data analysis has to consider at least two data generation components: the economic behavior model describing the stimulus–response behavior of the economic actors and the measurement model describing how and when this behavior was recorded and summarized. While it is usual to focus our attention on the former, a complete analysis must consider them both.

Four decades later, measurement is still routinely treated as an innocuous exercise in data preprocessing. Composite indices are constructed using standard recipes—unit weights, principal components, or factor scores—and inserted into regressions with little discussion of the formal assumptions

[1] See, for example, Spearman (1904) and Thurstone (1931) on factor models, Pearson (1901) and Hotelling (1933) on principal component analysis, Wright (1921) and Jöreskog (1970) on path analysis and structural equation modeling, and Jöreskog and Goldberger (1975) on multiple indicators multiple causes models.

[2] Composite indices are also used for non-regression purposes such as rankings of "development," "health," or "welfare." The Human Development Index is a well-known example; Ravallion (2012) provides a thorough discussion. Here, we focus on composite indices as covariates in regression analysis.

these choices require. The prevailing intuition suggests that aggregating multiple proxies naturally mitigates measurement error. We show that this intuition is often misplaced.

Our primary objective is to assess the econometric consequences of relying on such indices—both theoretically and empirically—and to provide practical guidance for researchers. We argue that index choice is an overlooked but critical identification choice despite the frequency with which such indices appear in economics and other disciplines. Researchers must think carefully and be transparent about the assumed relationship between observed manifest variables and the unobserved latent construct. This relationship—the measurement model—is an integral part of identification and must be made explicit. A central theme of the paper is that failure to discipline measurement choices cannot be remedied *ex post* by conventional robustness checks.

We are not the first to warn researchers of the perils of using proxies in general and *ad hoc* indices in particular. Concerns about the arbitrary nature of index construction have long been emphasized (e.g., Mundlak 1961; Samuelson 1983; Ravallion 2012). A related literature highlights how the inclusion of proxies generates residual confounding—biasing estimates of other model parameters through incomplete conditioning on the latent construct, even when the remainder of the model is correctly specified (e.g., Bollinger 2003; Erickson and Whited 2006; Bollinger and Minier 2015). Other contributions point to shortcomings of specific dimensionality-reduction techniques, including the use of principal components for structural estimation (Lubotsky and Wittenberg 2006; Krishnakumar and Nagar 2008; Greco et al. 2019; Mazziotta and Pareto 2019). Despite these repeated warnings, their implications have not been fully absorbed by applied researchers. Yet, as Mazziotta and Pareto (2016, p. 985) emphasize, "the choices of the researcher assume a fundamental role" and may lead to large biases and erroneous policy conclusions. Our contribution is to turn a diffuse warning into a structured framework and a clear path forward. By framing the problem through explicit measurement models—reflective and formative—and by discussing several easy-to-implement alternatives, we aim to improve the credibility and interpretability of empirical research.

Our paper is most closely related to Jarvis et al. (2003), Blundell et al. (2023), and Stoetzer et al. (2025). Jarvis et al. (2003) study similar issues in marketing research, emphasizing the lack of attention paid to measurement model choice, construct validity, and the consequences of misspecification. Although written more than two decades ago and in a different discipline, their critique applies directly to much contemporary empirical work in economics. Blundell et al. (2023) examine bias arising from the use of objective and subjective measures of health to proxy for latent health and propose an instrumental variables (IV) strategy valid in their context. Our analysis differs in that we do not rely on context-specific instruments and instead characterize what can and cannot be achieved using linear indices under different assumptions. Stoetzer et al. (2025) address closely related concerns in a setting where the latent construct is the *outcome* and the object of interest is the average treatment effect of an intervention. They show that two-step approaches—where the latent outcome is first estimated using principal component analysis (PCA) or related methods and the treatment effect is then expressed in standard deviations of the estimated latent construct—are generally biased. The authors propose a hierarchical item response theory model as an alternative. Our analysis complements their discussion by focusing on settings where the latent construct enters as a covariate and where its coefficient may or may not be the primary object of interest.

Our work also relates to Black and Smith (2006), Filmer and Scott (2012), and Knaus et al. (2020). Black and Smith (2006) construct factor-based indices of college quality using a small number of manifest variables and show how to consistently estimate the coefficient of interest under the assumption that measurement errors are uncorrelated across indicators. We relax this assumption, which is unlikely to hold in many applications. Filmer and Scott (2012) compare several methods for constructing household asset indices, focusing on rankings and descriptive properties rather than regression-based inference. Although addressing a different question, their comparative approach is closely aligned with ours. Similarly, Knaus et al. (2020) conduct an extensive simulation study comparing multiple estimators across a wide range of data-generating processes. In the same spirit, we compare the implications of multiple approaches to index construction on estimation and inference in linear regression models.

More broadly, our contribution relates to but is distinct from several strands of the literature. A long-standing body of work emphasizes the econometric consequences of measurement error when theoretical variables are imperfectly observed, highlighting how misspecification of the measurement process can fundamentally alter identification and inference (Griliches 1985). More recent work considers settings with multiple proxies for a latent construct, often motivated by the idea that combining information across proxies may improve empirical performance. A prominent example is Lubotsky and Wittenberg (2006), who derive an optimal linear combination of proxies under a single-factor structure, along with subsequent extensions that formalize alternative weighting and estimation strategies (e.g., Yang et al. 2023). In parallel, applied researchers routinely rely on dimensionality-reduction techniques such as PCA, exploratory factor analysis (EFA), or partial least squares (PLS), drawing on traditions from psychometrics and multivariate statistics. Our paper differs in emphasis and scope. Rather than proposing a single preferred estimator or weighting scheme, we ask a more basic question: what can and cannot be achieved when a latent regressor is replaced by a linear index in an otherwise correctly specified regression model, and what insights should guide applied work in light of these limits?

To address this question, we analyze the impact of several popular index creation techniques—including PCA, EFA, PLS, mean z-scores, and unit (or equal) weights—on the estimates of a multiple linear regression model.[3] The researcher may care about the coefficient on the index itself, or this coefficient may be a nuisance parameter with primary interest focused on other covariates.

Our analysis delivers four takeaways that we view as central for applied researchers. First, we show that index creation methods introduce nonclassical measurement error, with the nature of the error differing across methods and measurement models. Second, replacing a latent regressor with a composite index is not a neutral data-reduction step. Even when all proxies are well measured and substantively relevant, index construction embeds identifying assumptions, so *index choice is an identification choice*. Third, even within the class of linear indices, there is no universally valid solution. We establish a (near) impossibility result: no linear proxy index can guarantee consistent OLS estimates of all parameters in a multiple-regression model. Any index necessarily trades off approximation error in the latent construct against induced correlation with other regressors, making optimality inherently context- and estimand-specific. Finally, the consequences of proxy indexing extend beyond the coefficient on the latent construct

[3] We discuss non-index methods in Section 5.

itself. When covariates are correlated with the index or latent construct, incomplete conditioning on the latent variable induces residual confounding that can substantially distort estimation and inference *of all parameters in the model*, with the magnitude and direction of the bias and *even the sign* of the estimates depending on the choices made during index construction.

These results have direct implications for empirical practice. Credible applied work must treat the measurement model as a first-order design choice and adopt explicit principles governing index construction, transparency, and sensitivity analysis. Through theoretical analysis and simulations, we show that failure to do so can lead to unstable estimates and misleading inference even in otherwise carefully designed studies. Through a replication of Ortoleva and Snowberg (2015), who examine the effect of media exposure on overconfidence and overconfidence on political ideology, and an original study of local public health on changes in Republican vote share from 2020 to 2024, we demonstrate that these concerns can be empirically relevant and substantively important.

The remainder of the paper proceeds as follows. Section 2 formalizes our framework, laying out reflective-indicator and formative-indicator data structures, defining the class of linear indices, and characterizing how replacing the latent regressor x^* with a linear proxy index induces bias in OLS. We also discuss the optimal linear index, which leads to a stark impossibility result: no linear index exists such that the OLS estimates of all parameters in a multiple-regression model are guaranteed to be consistent. Sections 3 and 4 examine specific index constructions as illustrations of these general principles. Section 5 discusses alternatives to index creation, including the approach proposed by Lubotsky and Wittenberg (2006), our extension based on Yang et al. (2023), and related IV and structural strategies. Section 6 presents simulation evidence tailored to applied settings. Section 7 distills the theoretical and simulation evidence into practical principles for applied researchers, emphasizing transparency about measurement models and systematic sensitivity analysis across credible index constructions and non-index alternatives. Sections 8 and 9 illustrate the stakes empirically, first by revisiting Ortoleva and Snowberg (2015) and then through an original analysis linking the Congressional District Health Dashboard to precinct-aggregated presidential returns. Section 10 concludes.

2 Index Creation

2.1 Setup

We focus on linear multiple regression models of the form

$$y = \alpha + \beta x^* + \gamma w + \varepsilon, \tag{1}$$

where y is the outcome, x^* is the latent construct, w is an observed covariate, and ε is the error term. Including only a single additional covariate w is without loss of generality: other observed determinants can be considered as partialled out, with x^* and w denoting residuals from linear projections on remaining covariates. The research objective may be to estimate β, or β may be a nuisance parameter with primary interest in γ. If β is of interest, x^* has no intrinsic scale, so the meaning of β beyond its sign is unclear. To address this, we conceptualize x^* as standardized, so β represents the marginal effect of a one standard deviation increase in the latent variable.

We assume all standard assumptions of the Classical Linear Regression Model hold; OLS estimation of Equation (1) produces unbiased and consistent estimates if x^* is observed. However, instead of observing x^*, there exist $\mathcal{J}$ manifest variables, $z_1, z_2, \ldots, z_{\mathcal{J}}$, related to x^*. Without loss of generality, we assume researchers observe the first $J \leq \mathcal{J}$ variables.[4]

An index $x := g(z_1, z_2, \ldots, z_J)$ is any scalar aggregate of the J observed manifest variables intended to proxy for x^*. Linear indices are most common in practice.[5] These take the form:

$$x := \sum\nolimits_j \lambda_j \left(\frac{z_j - \bar{z}_j}{\sigma_{z_j}} \right) \tag{2}$$

if the manifest variables are standardized, where $\bar{z}_j$ and σ_{z_j} denote the sample mean and standard deviation of z_j, and

$$x := \sum\nolimits_j \lambda_j z_j \tag{3}$$

if not standardized. The weights λ_j must be chosen and need not be positive or sum to one.

Index creation involves four steps (Mazziotta and Pareto 2016): (i) defining the latent construct, (ii) selecting manifest variables, (iii) transforming (or not) the manifest variables, and (iv) choosing the aggregation method. A key part of step one is specifying a "well-defined measurement model" (Mazziotta and Pareto 2019, p. 452). Researchers often gloss over the fact that "constructing a composite index ... is a conceptual, as well as mathematical, operation" (Mazziotta and Pareto 2019, p. 452).

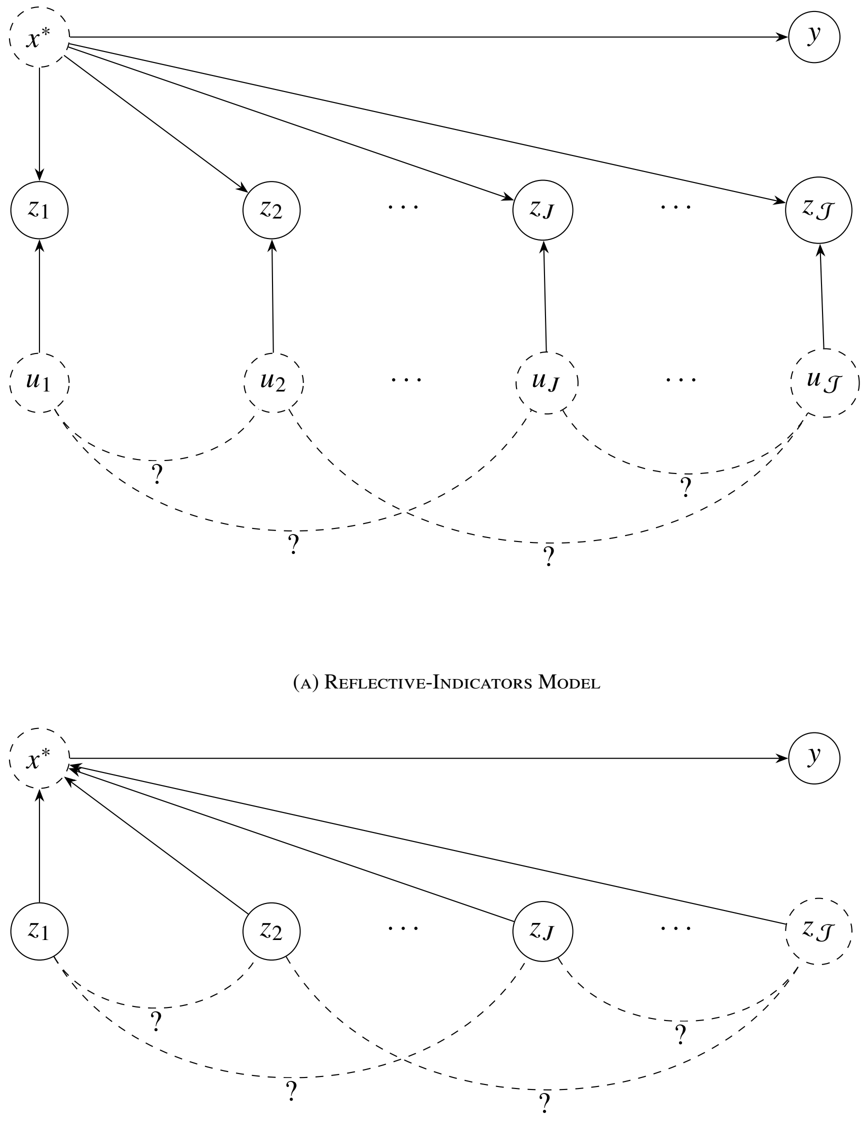
Figure 1 illustrates two popular measurement models (Jarvis et al. 2003).[6] Panel A displays the *reflective-indicators model*, where manifest variables "reflect" x^* and other unmodeled attributes u. Panel B presents the *formative-indicators model*, where manifest variables "form" x^*. Jarvis et al. (2003) discuss four contrasts to guide model selection. First, causation flows from x^* to manifest variables in the reflective model; the reverse in the formative model. Second, manifest variables are interchangeable in the reflective model but not in the formative model. Third, manifest variables are necessarily correlated in the reflective model but need not be in the formative model. Fourth, manifest variables should have the same antecedents and consequences in the reflective model but not necessarily in the formative model.

These differences have implications for measurement error. In the reflective model, measurement error arises from u; each manifest variable is an error-ridden proxy for x^*. Observing $J < \mathcal{J}$ exacerbates this if more manifest variables produce a better approximation. In the formative model, measurement error arises if the aggregation scheme mapping manifest variables into x^* is incorrect, or due to unobserved manifest variables z_j, $j = J + 1, ..., \mathcal{J}$. Conceptually, omitting these variables is equivalent to using incorrect weights by assigning them zero weight.

[4] Schennach (2020, p. 496) defines manifest variables as "related to the true value of the variable of interest but may be expressed in different units or may even be nonlinearly related to the true value," noting they are "formally equivalent to measurements with general nonclassical error." We use the term 'manifest variables' throughout.

[5] There is no theoretical reason to restrict attention to linear indices, but applied researchers nearly always do (see, e.g., Black and Smith 2006; Greco et al. 2019; Blundell et al. 2023). Mazziotta and Pareto (2016) and Mazziotta and Pareto (2018) discuss implications of linearity and alternatives.

[6] Meijer et al. (2025) distinguishes three types: (i) manifest variables affected by the latent variable, (ii) manifest variables affecting the latent variable, and (iii) manifest variables correlated with the latent variable but with no causal relationship. We focus on the first two.



(a) Reflective-Indicators Model

(b) Formative-Indicators Model

Figure 1
Two Common Measurement Models

Notes.— Directed Acyclic Graphs depicting the relationship between manifest variables z_j, $j = 1, ..., \mathcal{J}$ and x^*. Dashed circles denote unobserved variables. Dashed lines indicate possible relationships. Other observed determinants of y (w in Equation (1)) are omitted for simplicity.

Millimet et al. (2018) is an example of the reflective model: latent financial literacy drives the answers to observed financial questions, along with other unobserved factors such as survey effort or luck guessing, which in turn are used to create an index of financial literacy. Maccini and Yang (2009) and Filmer and Scott (2012) are examples of the formative model, using observed assets to construct household asset indices where latent assets is an aggregate of all assets, observed and unobserved. Angelini et al. (2017) is another example, using reported life satisfaction across several domains to construct an index of overall life satisfaction. Blundell et al. (2023) consider both: the reflective model for self-reported subjective health and the formative model for objective health measures. As we confirm later, this distinction matters considerably in practice; however, to our knowledge, it has not appeared in the economics literature aside from Blundell et al. (2023).

Steps two through four are equally critical. Different choices of manifest variables and aggregation methods $g(\cdot)$ generally lead to different values of x and hence different levels of similarity between x and x^*. Mazziotta and Pareto (2016, p. 985) call aggregation the "delicate phase" of index creation. Greco et al. (2018, p. 586) similarly refer to it as "perhaps the most pernicious problem," noting that "the weights set is clearly the manifest problem for composite indices." For now, we take $\mathbf{z} := [z_1 \; z_2 \; \cdots \; z_J]$ as given and abstract from manifest variable selection.

2.1.1 Reflective-Indicators Model

We make the following assumption regarding the manifest variables under the reflective-indicators model.

Assumption 1 (*Manifest Variables*)**.** *Let (y_i, x_i^*, w_i) be iid draws from the distribution of (Y, X^*, W). The manifest variables are generated according to*

$$z_j = x^* + u_j, \quad j = 1, 2, \ldots, \mathcal{J}, \tag{4}$$

where $\mathtt{E}\left(u_j\right) = 0 \; \forall j$, $\mathtt{Var}\left(u_j\right) = \sigma_{u_j}^2 \; \forall j$, $\mathtt{Cov}\left(u_j, u_{j'}\right) = 0 \; \forall j \neq j'$, $\mathtt{Cov}\left(x^*, u_j\right) = 0 \; \forall j$, $\mathtt{Cov}\left(\varepsilon, u_j\right) = 0 \; \forall j$, $\mathtt{Cov}\left(w, u_j\right) = 0 \; \forall j$, *and* $\mathtt{Var}(x^*) = 1$.

Under Assumption 1, each z_j is an unbiased reflection of x^*, the z_j are independent conditional on x^* and independent of the regression error, and the u_j satisfy classical measurement error assumptions.

Much of Assumption 1 simplifies the exposition. In practice, several complications may arise. First, manifest variables often have their own location and scale, $z_j = \omega_{0j} + \omega_{1j} x^* + u_j$. Allowing $\omega_{1j} \neq 1$ complicates interpretation since manifest variable units differ from the intrinsic units of x^*. The typical solution is to normalize $\omega_{1j} = 1$ for a particular j, allowing x^* to be interpreted in that variable's units (e.g., Agostinelli and Wiswall 2025). Here, we omit location and scale parameters and instead normalize $\mathtt{Var}(x^*) = 1$. Second, measurement errors are likely correlated across manifest variables, $\mathtt{Cov}\left(u_j, u_{j'}\right) \neq 0$ for some $j \neq j'$ (e.g., Lubotsky and Wittenberg 2006). Third, measurement errors are often correlated with the truth, $\mathtt{Cov}\left(x^*, u_j\right) \neq 0$. Mean-reversion ($\mathtt{Cov}\left(x^*, u_j\right) < 0$) has been documented in self-reported earnings (Millimet and Parmeter 2025), body mass index (Dong and Millimet 2024), life satisfaction (Angelini et al. 2017), grade point averages (Kuncel et al. 2005), and more. We relax some of these assumptions in Section 6.

The index or proxy error is

$$\mu := x - x^* = \begin{cases} \left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right) x^* - \left(\sum_j \frac{\lambda_j}{\sigma_{z_j}}\right) \bar{x}^* + \sum_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j) & \text{(standardized)} \\ \left(\sum_j \lambda_j - 1\right) x^* + \sum_j \lambda_j u_j & \text{(non-standardized)} \end{cases} \tag{5}$$

Critically, even when errors in Equation (4) are classical, the proxy error μ is unlikely to be. This is summarized in the following proposition.

Proposition 1. *Under Assumption 1, the proxy error has the following asymptotic properties:*

Quantity	Standardized Case	Non-Standardized Case
$\mathtt{E}[\mu]$	$-\mathtt{E}[x^*]$	$\left(\sum_j \lambda_j - 1\right) \mathtt{E}[x^*]$
$\mathtt{Var}(\mu)$	$\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right)^2 + \sum_j \left(\frac{\lambda_j}{\sigma_{z_j}}\right)^2 \sigma^2_{u_j}$	$\left(\sum_j \lambda_j - 1\right)^2 + \sum_j \lambda_j^2 \sigma^2_{u_j}$
$\mathtt{Cov}(x^*, \mu)$	$\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1$	$\sum_j \lambda_j - 1$
$\mathtt{Cov}(w, \mu)$	$\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right) \mathtt{Cov}(x^*, w)$	$\left(\sum_j \lambda_j - 1\right) \mathtt{Cov}(x^*, w)$

In the standardized case, the proxy error is classical measurement error if and only if x^* is mean zero and $\sum_j \lambda_j/\sigma_{z_j} = 1$. In the non-standardized case, the proxy error is classical measurement error if and only if $\sum_j \lambda_j = 1$. In addition, $\mathtt{Cov}(w, \mu) = 0$ if and only if either (i) $\sum_j \lambda_j/\sigma_{z_j} = 1$ or (ii) $\mathtt{Cov}(x^*, w) = 0$ (with λ_j replacing λ_j/σ_{z_j} in the non-standardized case).
Proof: See Appendix A. ■

2.1.2 Formative-Indicators Model

For the formative-indicators model, the manifest variables are taken as given. The following assumption is made concerning x^*.

Assumption 2 (*Latent x^**)**.** *Let $(y_i, \{z_{ji}\}_{j=1}^{\mathcal{J}}, w_i)$ be iid draws from the distribution of $(Y, \{Z_j\}_{j=1}^{\mathcal{J}}, W)$. The latent construct x^* is generated according to*

$$x^* = \sum_{j=1}^{\mathcal{J}} \lambda_j^* z_j, \tag{6}$$

where $\lambda_j^* \neq 0 \; \forall j$ and are normalized such that $\mathtt{Var}(x^*) = 1$.

Under Assumption 2, x^* is a weighted sum of all manifest variables ($j = 1, ..., \mathcal{J}$) with weights λ_j^*. The variance normalization facilitates interpretation of β. Alternatively, one could normalize $\lambda_j^* = 1$ for some j, allowing x^* to be interpreted in that variable's units. The linear index x is still given by Equation (2) if standardized and Equation (3) otherwise.

An important scope condition applies. Our framework presumes that x^* is a latent construct distinct from the index x—that is, the "true" aggregation exists independently of the researcher's measurement choices. In many policy settings, however, the construct *is* the index by definition: a specific poverty score, a regulatory compliance measure, or a composite ranking may be defined by a particular set of weights established by convention or statute. In such "index-as-definition" cases, there is no latent x^* apart from the chosen aggregate, so deviations between x and x^* are not measurement error but rather reflect a different construct entirely. In this case, the relevant concern is not measurement error but whether the defined index captures a economically coherent quantity (see, e.g., Ravallion 2012). In contrast, our warnings about proxy error apply when the researcher believes a latent truth exists and the index is an imperfect attempt to recover it.

In the formative-indicators model, the index or proxy error is

$$\mu := x - x^* = \begin{cases} \sum_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) z_j - \sum_j \frac{\lambda_j}{\sigma_{z_j}} \bar{z}_j & \text{(standardized)} \\ \sum_j \left(\lambda_j - \lambda_j^*\right) z_j & \text{(non-standardized)} \end{cases} \tag{7}$$

As in the reflective model, μ is unlikely to be classical measurement error.[7] This is summarized in the following proposition.

Proposition 2. *Under Assumption 2, the proxy error has the following properties:*

Quantity	**Standardized Case**	**Non-Standardized Case**
$\mathtt{E}[\mu]$	$-\mathtt{E}[x^*]$	$\sum_j \lambda_j \mathtt{E}[z_j] - \mathtt{E}[x^*]$
$\mathtt{Var}\,(\mu)$	$\sum_j \sum_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right)\left(\frac{\lambda_k}{\sigma_{z_k}} - \lambda_k^*\right) \mathtt{Cov}(z_j, z_k)$	$\sum_j \sum_k \left(\lambda_j - \lambda_j^*\right)\left(\lambda_k - \lambda_k^*\right) \mathtt{Cov}(z_j, z_k)$
$\mathtt{Cov}\,(x^*, \mu)$	$\sum_j \sum_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) \lambda_k^* \mathtt{Cov}(z_j, z_k)$	$\sum_j \sum_k \left(\lambda_j - \lambda_j^*\right) \lambda_k^* \mathtt{Cov}(z_j, z_k)$
$\mathtt{Cov}\,(w, \mu)$	$\sum_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) \mathtt{Cov}(z_j, w)$	$\sum_j \left(\lambda_j - \lambda_j^*\right) \mathtt{Cov}(z_j, w)$

In the standardized case, the proxy error is classical measurement error if and only if x^* is mean zero and $\left(\lambda_j/\sigma_{z_j}\right) = \lambda_j^* \; \forall j$. In the non-standardized case, the proxy error is classical measurement error if and only if $\lambda_j = \lambda_j^* \; \forall j$, i.e., $\mu \equiv 0$ (the degenerate no-error case). In addition, $\mathtt{Cov}(w, \mu)$ *is zero if and only if* $\mathtt{Cov}(z_j, w) = 0 \; \forall j$ *(excluding knife-edge cases).*
Proof: See Appendix B. ■

2.2 Properties of OLS

To assess the impact of using an proxy index in lieu of x^*, we make the following assumption regarding the structural data-generating process.

Assumption 3 (*Data-Generating Process (DGP)*)**.** *The outcome y is generated according to Equation* (1) *with* $\mathtt{E}[\varepsilon] = 0$ *and* $\mathtt{E}\left[\widetilde{\mathbf{w}}\varepsilon\right] = 0$*, where* $\widetilde{\mathbf{w}} := [x^* \; w]$.

Under Assumption 3, OLS provides unbiased and consistent estimates if x^* is observed. Thus, proxy error is the sole source of misspecification. This yields the following result.

[7] The formative model resembles the Berkson measurement error model (e.g., Meijer et al. 2025) where the latent variable is a linear function of observed proxies. However, here not all proxies are observed and they may be correlated.

Proposition 3. *Under Assumption 3, either Assumption 1 or 2, and replacing x^* with x in Equation (1), the OLS estimates of β and γ converge to*

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta\left[\frac{\delta_x^* - \frac{\delta^*}{\delta}R^2_{x|w}}{1 - R^2_{x|w}}\right] \tag{8}$$

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R^2_{x|w}}\right] = \gamma + \delta\left[\beta\frac{\delta^*}{\delta} - \mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}}\right] \tag{9}$$

where $\delta^* := \mathtt{Cov}(x^*, w)/\mathtt{Var}(w)$ is the coefficient from regressing x^* on w, $\delta := \mathtt{Cov}(x, w)/\mathtt{Var}(w)$ is the coefficient from regressing x on w, $\delta_x^* := \mathtt{Cov}(x^*, x)/\mathtt{Var}(x)$ is the coefficient from regressing x^* on x, and $R^2_{x|w}$ is the R^2 from regressing x on w. This holds regardless of whether the data structure conforms to the reflective or formative model and regardless of whether x uses standardized or non-standardized manifest variables. If $\mathtt{Cov}(x^*, w) = 0$ or, in the formative model, $\mathtt{Cov}(z_j, w) = 0\ \forall j$, then $\delta^* = \delta = 0$ and this simplifies to

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta\delta_x^* \tag{10}$$

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma. \tag{11}$$

Proof: See Appendix C.1. ■

Proposition 3 shows that OLS does not necessarily suffer from attenuation bias and may even have the wrong sign. In Equation (8), δ_x^* captures the slope of the regression of x^* on x and can exceed or fall below one; $\delta^*/\delta = \mathtt{Cov}(x^*, w)/\mathtt{Cov}(x, w)$ equals one in absolute value if μ is classical measurement error but otherwise may differ. Nothing precludes the bracketed term in Equation (8) from being negative. Rearranging, the numerator equals $\mathtt{Cov}(x^*, x|w)/\mathtt{Var}(x)$, implying sign reversal when x^* and x are negatively correlated after partialling out w.[8]

Proposition 3 also shows that γ will be biased unless $\mathtt{Cov}(x^*, w) = 0$ (reflective) or $\mathtt{Cov}(z_j, w) = 0\ \forall j$ (formative) (see, e.g., Bollinger 2003; Bollinger and Minier 2015). The mechanism is *residual confounding*: because x is an imperfect proxy for x^*, it fails to absorb all latent variation, leaving w correlated with the composite error $\varepsilon - \beta\mu$. Researchers must be cautious when using proxies even if the proxy is not of interest; generally, *all coefficients are biased in unknown direction*. The magnitude depends on the relationship between the proxy index and w and the bias in the estimate of the β.

Lastly, it is crucial to note that in the reflective model, the IV estimator—using $\mathbf{z}_{-j}$ ($\mathbf{z}$ omitting z_j) as instruments for z_j—is consistent for β and γ under Assumption 1. However, this IV procedure is inconsistent if measurement errors are correlated across manifest variables ($\mathtt{Cov}(u_j, u_{j'}) \neq 0$ for some $j \neq j'$).[9] As correlated measurement errors are likely the norm, this is not a reliable solution. Moreover,

[8] To see this, note that

$$\begin{aligned}\delta_x^* - \frac{\delta^*}{\delta}R^2_{x|w} &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} - \left[\frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)}\right]\left[\frac{\mathtt{Cov}(x, w)^2}{\mathtt{Var}(x)\mathtt{Var}(w)}\right] \\ &= \frac{1}{\mathtt{Var}(x)}\left[\mathtt{Cov}(x^*, x) - \frac{\mathtt{Cov}(x^*, w)\mathtt{Cov}(x, w)}{\mathtt{Var}(w)}\right] = \frac{\mathtt{Cov}(x^*, x|w)}{\mathtt{Var}(x)}.\end{aligned}$$

[9] With correlated measurement errors, the instruments $\mathbf{z}_{-j}$ are correlated with the composite error.

even with uncorrelated measurement errors, the IV estimate of γ is inconsistent if w is correlated with μ (unless $\mathbf{z}_{-j}$ provides sufficient exclusion restrictions for x and w). Finally, this IV strategy is *never* valid in the formative model: since x^* depends on all $\mathcal{J}$ manifest variables by Equation (6), an index formed from a subset necessarily has measurement error equal to the omitted variables' aggregation, making excluded manifest variables invalid instruments. The failure of IV except under strong assumptions makes it critical to consider different index choices and alternative estimators.

2.3 Optimal Linear Index

Before turning to specific indices, we ask whether an "optimal" linear index exists. Consider a generic linear index $x = \sum_{j=1}^{J} \lambda_j h(z_j)$, where $h(z_j)$ equals z_j or its standardized version. Define the optimal index as the one minimizing the absolute asymptotic bias in Equation (8) or (9). This leads to the following result.

Proposition 4. *Under Assumption 3 and either Assumption 1 or 2 and replacing x^* with x in Equation (1), the OLS estimate of β is consistent if*

$$\texttt{Cov}(x, \mu \mid w) = 0.$$

The OLS estimate of γ is consistent if

$$\texttt{Cov}(w, \mu \mid x) = 0.$$

Proof: See Appendix C.2. ■

Proposition 4 provides critical insights. First, x and w both must be uncorrelated with the proxy error conditional on the other for consistent estimation of the parameters. Since $\texttt{Cov}(x, \mu \mid w) = \texttt{Cov}(x^*, \mu \mid w) + \texttt{Var}(\mu \mid w)$, consistency requires $\texttt{Cov}(x^*, \mu \mid w) < 0$: the proxy error must be mean-reverting conditional on w. Second, and most importantly, simultaneously obtaining consistent estimates of β and γ is very unlikely in practice, as shown in the following corollary.

Corollary 1. *Define the optimal linear index weights:*

$$x^o = \sum\nolimits_j \lambda_j^o z_j. \tag{12}$$

Under Assumption 3 and either Assumption 1 or 2, assuming $\beta \neq 0$ and $\texttt{Var}(w|\mathbf{z}) > 0$, and replacing x^* with x in Equation (1), the OLS estimate of β is consistent if $\lambda_j^o = \delta_j^*$, where δ_j^* are coefficients from the linear projection of x^* on observed variables:

$$x^* = \sum\nolimits_j \delta_j^* z_j + \zeta w + \nu. \tag{13}$$

However, simultaneous consistency of $\widehat{\beta}$ and $\widehat{\gamma}$ requires (a) the absence of a structural link between the covariate and the latent variable ($\zeta = 0$), or (b) mean-reverting measurement error such that $\texttt{Cov}(x^*, \mu) = -\texttt{Var}(\mu)$.
Proof: See Appendix C.2 and C.3. ■

Corollary 1 is quite telling. First, since x^* is unobserved, δ_j^* is not estimable and no feasible linear index can be obtained. Second, even if δ_j^* were known, the OLS estimates of the remaining parameters are generally inconsistent. Simultaneous consistency of $\widehat{\beta}$ and $\widehat{\gamma}$ requires the absence of a structural link between the covariate and the latent variable ($\zeta = 0$). This is unlikely to hold in most applications where w is not a randomized treatment. In the presence of such a link ($\zeta \neq 0$), consistency can only be achieved through a specific, nonclassical form of mean-reverting measurement error that exactly offsets the structural contamination—a condition that is generally incompatible with the construction of a linear index. This (near) *impossibility* result for γ should alarm researchers as it brings to light a trade-off between the consistency of the two coefficients.

3 The Case Against PCA Indices

3.1 PCA for Index Creation

Perhaps the most popular approach to index construction—particularly in economics—is PCA (Pearson 1901; Hotelling 1933).[10] PCA transforms J manifest variables into J orthogonal linear combinations (principal components) given by

$$x^{\mathtt{PCA}(k)} := \sum\nolimits_j v_j^{(k)} z_j, \tag{14}$$

where k indexes the principal components ($k = 1, ..., J$) and $v_j^{(k)}$ are the weights for the $k^{\mathtt{th}}$ component. The weights maximize the variance of $x^{\mathtt{PCA}(k)}$ subject to orthogonality to $x^{\mathtt{PCA}(k')}$, $k' = 1, ..., k-1$.

Since researchers typically rely on the first principal component as the index, we focus on PC1 throughout. Formally, the PCA index is $x^{\mathtt{PCA}} := \sum_j v_j z_j$ (dropping the superscript), where the weight vector $\mathbf{v} = [v_1 \; v_2 \; \cdots \; v_J]'$ solves

$$\mathbf{v}^{\mathtt{PCA}} = \arg\max_{\|\mathbf{v}\|=1} \mathrm{Var}(\mathbf{v}'\mathbf{z}) = \arg\max_{\|\mathbf{v}\|=1} \mathbf{v}'\Sigma_{\mathbf{zz}}\mathbf{v} \tag{15}$$

and $\Sigma_{\mathbf{zz}}$ is the $(J \times J)$ covariance matrix of the manifest variables. The solution is the eigenvector corresponding to the largest eigenvalue of $\Sigma_{\mathbf{zz}}$, with the unit-norm constraint $\|\mathbf{v}\| = 1$ preventing arbitrary scaling. The resulting index has variance equal to this largest eigenvalue.

Researchers typically standardize manifest variables to have zero mean and unit variance before applying PCA. Standardization can be particularly important when variables are measured in different units. With standardized data, $\mathbf{v}^{\mathtt{PCA}}$ becomes the eigenvector corresponding to the largest eigenvalue of the correlation matrix.

The widespread adoption of PCA stems from several attractive properties. First, it provides effective dimensionality reduction by compressing J manifest variables into a scalar capturing maximal joint variation. Second, the weights v_j are determined objectively from the data, avoiding arbitrary weighting schemes (Jolliffe and Cadima 2016). Variables exhibiting greater covariation receive larger weights in absolute value, often interpreted as revealing relative importance. Third, when multiple components are

[10] A search on Google Scholar for "principal component analysis" AND "index" AND "regression" yields 717,000 hits; 29,900 since 2024. Accessed 2 December 2025. A search on Paul Goldsmith-Pinkham's website indicates that roughly 1.5% of papers published in the *American Economic Review* and four *American Economic Journal* journals since 2010 contain "principal component analysis" AND "index" (67 papers), with an upward trend since 2017. According to Dobriban (2020, p. 2824), "Factor Analysis (FA) and Principal Component Analysis (PCA), the unsupervised discovery of components governing variation in the data, is performed routinely in thousands of studies every year."

used, they are orthogonal by construction, eliminating multicollinearity concerns.

Yet from an econometric standpoint, PCA has several limitations.[11] First, and most critically, PCA ignores the regression context entirely. The optimization in Equation (15) depends solely on $\Sigma_{\mathbf{zz}}$ and incorporates no information about the relationship between manifest variables and either the outcome y or the latent construct x^*. PCA is an unsupervised technique that reduces dimensionality without reference to the dependent variable or structural relationships—a critical shortcoming when the goal is estimation or inference.

This has a crucial implication: PCA treats all variation in manifest variables as equally valuable, regardless of whether it stems from the common latent factor (signal) or idiosyncratic noise. If nuisance variance dominates the covariance structure, PCA produces a poor index for regression. As Mazziotta and Pareto (2019) argue, there is a fundamental distinction between dimensionality reduction—which PCA is designed for—and index construction for regression—which it is not.

Second, the unit-norm constraint introduces systematic measurement error with problematic properties. The normalization $\sum_j v_j^2 = 1$ alters the measurement error structure in ways that depend on the underlying data-generating process. In the reflective model with non-standardized variables, Proposition 1 shows this constraint is incompatible with classical measurement error, which requires the unit-sum constraint $\sum_j v_j = 1$. With standardized variables, classical measurement error requires $\sum_j v_j/\sigma_{z_j} = 1$, a condition unlikely to be satisfied by PCA weights. In the formative model, any deviation between PCA weights and the true structural weights generates nonclassical measurement error. While nonclassical measurement error is not inherently more problematic than classical error, the latter provides clear guidance about bias direction, facilitating interpretation and potential correction. The nonclassical error introduced by PCA can bias estimates in unpredictable directions and magnitudes.

Third, the scaling of the PCA index may differ from that of x^*, leading to estimates that are too large or too small relative to β (Black and Smith 2006). Finally, even if the researcher has no interest in β—the index is included as a control—measurement error will bias estimates of γ if x and w are correlated (see, e.g., Hanushek and Jackson 1977; Griliches 1986; Bollen 1989).[12]

3.2 Properties of PCA Indices

In the reflective model with non-standardized variables, PCA weights favor manifest variables with larger total variance, meaning noisier variables (larger $\sigma^2_{u_j}$) receive disproportionate weight. Standardization reduces this tendency as the weights become the leading eigenvector of the correlation matrix. In the formative model, PCA is more problematic because $\Sigma_{\mathbf{zz}}$ is not determined by a common latent factor; the z_j's can have arbitrary covariances. Thus, PCA is relatively better suited under the reflective model with standardized manifest variables (Mazziotta and Pareto 2019).[13]

To illustrate, consider the reflective model with two manifest variables measuring the same x^* with different noise levels: $z_1 = x^* + u_1$ with $\mathtt{Var}(u_1) = \sigma^2_{u_1}$ and $z_2 = x^* + u_2$ with $\mathtt{Var}(u_2) = \sigma^2_{u_2}$, assuming

[11] Mazziotta and Pareto (2019) provide detailed discussion of PCA's pros and cons for index construction.

[12] See Hünermund et al. (2025) for discussion on control variable choice. Zhang and Lee (2025) characterize when including a mismeasured covariate is preferred over exclusion when γ is an average treatment effect.

[13] Positing a formative model for objective health measures, Blundell et al. (2023, p. 291) state: "While it would also be possible to construct an index of health based on the objective variables, it would not be as compelling to do so as objective measures reflect different aspects of health rather than the same latent index."

$\texttt{Cov}(u_1, u_2) = 0$. If $\sigma^2_{u_2} \gg \sigma^2_{u_1}$, then z_2 exhibits greater total variance and PCA gives it more weight. In the extreme, $x^{\texttt{PCA}} \approx z_2$, inheriting z_2's low reliability and severely biasing OLS. Concretely, let $\texttt{Var}(x^*) = 1$, $\sigma^2_{u_1} = 0.5$ and $\sigma^2_{u_2} = 4$. The PCA weights are $v_1^{\texttt{PCA}} \approx 0.26$ and $v_2^{\texttt{PCA}} \approx 0.97$ if non-standardized; $v_1^{\texttt{PCA}} = v_2^{\texttt{PCA}} \approx 0.71$ if standardized, implying weights on the non-standardized manifest variables of approximately 0.58 and 0.32 (see Appendix D).[14] The optimal (infeasible) weights obtained from the linear projection of x^* on z_1 and z_2 are $\lambda_1^o \approx 0.62$ and $\lambda_2^o \approx 0.08$.[15] Thus, PCA puts more weight on the noisier indicator than optimal, although the problem is less severe with standardization.

3.3 OLS Bias When Using PCA Indices

We now quantify the OLS bias when using a PCA index. The following proposition combines Proposition 3 with the properties of $x^{\texttt{PCA}}$ in the reflective model.

Proposition 5. *Under Assumptions 1 and 3 and replacing x^* with $x^{\texttt{PCA}}$ in Equation (1), the OLS estimates converge to*

$$\texttt{plim}\, \widehat{\beta}^{\texttt{PCA}} = \beta \left[\frac{1}{\sum_j c_j} \cdot \frac{\sigma - R^2_{x|w}}{1 - R^2_{x|w}} \right]$$

and

$$\texttt{plim}\, \widehat{\gamma}^{\texttt{PCA}} = \gamma + \beta\delta^* \left[\frac{1-\sigma}{1 - R^2_{x|w}} \right],$$

where $\sigma := (\sum_j c_j)^2 / \texttt{Var}(x^{\texttt{PCA}})$ is the ratio of the signal variance to total variance in the index, and $c_j := v_j / \sigma_{z_j}$ if the manifest variables are standardized and $c_j := v_j$ if not.
Proof: See Appendix D.5. ■

As shown in the appendix, $x^{\texttt{PCA}} = \left(\sum_j c_j\right) x^* + \psi$, where ψ captures terms not depending on x^*, and $\beta / \sum_j c_j$ is the rescaled slope coefficient arising from the different scales of $x^{\texttt{PCA}}$ and x^*. Since signal variance cannot exceed total variance, $\sigma \leq 1$, with equality only when all measurement errors have zero variance.[16] The scaling term has the following bounds:

$$\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} \in \left(\frac{1}{\sigma_{z_{j^*}}}, \sqrt{\sum\nolimits_j \frac{1}{\sigma^2_{z_j}}} \,\right] \qquad \text{and} \qquad \sum\nolimits_j v_j \in \left(1, \sqrt{J}\,\right]$$

where $\sigma_{z_{j^*}} = \max(\sigma_{z_1}, \ldots, \sigma_{z_J})$ (see Appendix D.4). These bounds rely on $v_j > 0\ \forall j$, which is guaranteed under the reflective model by the Perron–Frobenius theorem (since $\mathbf{\Sigma_{zz}}$ has strictly positive off-diagonal entries), but need not hold in the formative model where the covariance structure is unrestricted. The scaling term may be above or below one when standardized, while the scaling term is strictly larger than one in the non-standardized case (for $J \geq 2$). The term $(\sigma - R^2_{x|w})/(1 - R^2_{x|w})$ can be less than,

[14] For $J = 2$ (or when all pairwise correlations are equal), standardized PCA weights are equal; mapping back to the raw scale yields coefficients proportional to $1/\sigma_{z_j}$. With $J > 2$ and heterogeneous idiosyncratic variances, standardized PCA weights are generally unequal.

[15] Alternatively, maximizing the reliability ratio $\texttt{Var}(x^*)/\texttt{Var}(x)$ subject to $\sum_j \lambda_j = 1$ yields $\lambda_1^\star \approx 0.89$ and $\lambda_2^\star \approx 0.11$. However, since the measurement error is unlikely to be classical, maximizing the reliability ratio is not necessarily desirable.

[16] From the definition, $\sigma = (\sum_j c_j)^2 \texttt{Var}(x^*)/\texttt{Var}(x)$. In the reflective model, $\texttt{Var}(x) = (\sum_j c_j)^2 \texttt{Var}(x^*) + \sum_j c_j^2 \sigma^2_{u_j}$, so $\sigma = (\sum_j c_j)^2 / [(\sum_j c_j)^2 + \sum_j c_j^2 \sigma^2_{u_j}] \leq 1$.

equal to, or greater than one depending on the reliability of the index relative to the correlation between x and w. Thus, both attenuation and expansion bias are possible for $\widehat{\beta}$ (though not sign reversal, since $\sigma > 0$ and all terms preserve sign). The direction of bias in $\widehat{\gamma}$ depends on the sign of $\beta\delta^*$, and sign reversal is possible since the term in brackets can be positive or negative.

Under the formative model, we have the following proposition.

Proposition 6. *Under Assumptions 2 and 3 and replacing x^* with x^{PCA} in Equation (1), the OLS estimates converge to*

$$\texttt{plim}\,\widehat{\beta}^{\text{PCA}} = \beta \left[\frac{\frac{1}{\varphi}\sum_j \sum_k \lambda_j^* c_k \texttt{Cov}(z_j, z_k) - \left(\frac{\sum_j \lambda_j^* \texttt{Cov}(z_j, w)}{\sum_j c_j \texttt{Cov}(z_j, w)} \right) R^2_{x|w}}{1 - R^2_{x|w}} \right]$$

and

$$\texttt{plim}\,\widehat{\gamma}^{\text{PCA}} = \gamma + \beta\delta \left[\frac{\frac{\sum_j \lambda_j^* \texttt{Cov}(z_j, w)}{\sum_j c_j \texttt{Cov}(z_j, w)} - \frac{1}{\varphi}\sum_j \sum_k \lambda_j^* c_k \texttt{Cov}(z_j, z_k)}{1 - R^2_{x|w}} \right],$$

where φ is the largest eigenvalue of the relevant matrix, $\texttt{Var}(x^{\text{PCA}}) = \varphi$, $c_j := v_j/\sigma_{z_j}$ if standardized and $c_j := v_j$ otherwise, with $c_j = 0$ for unobserved manifest variables.
Proof: See Appendix D.5. ■

The plims depend on $\texttt{Cov}(x^*, x)$ and differences between true weights λ_j^* and PCA weights c_j scaled by the manifest variables' relationships with w. As in the reflective model, either estimate may suffer from attenuation or expansion bias. However, unlike the reflective model, in the formative model both estimates may experience sign reversal. Sign reversal for β requires that the numerator is negative, occurring if the first term is small (PCA weights misaligned, assigning large weights to unimportant manifest variables and vice versa) and/or the second term is large (PCA assigns small weights to manifest variables highly correlated with w).

3.4 Discussion

PCA does not generally solve or even minimize problems arising from measurement error when using a composite index in lieu of a latent construct. This is true even if all manifest variables are unbiased measures of x^* (reflective model) or inputs into x^* (formative model). Moreover, the biases of $\widehat{\beta}$ and $\widehat{\gamma}$ are not guaranteed to be toward zero and estimates may converge to the wrong sign.

It is useful to distinguish two categories of problems. The first is *recoverable*: the PCA index scale may differ from that of x^*, but this is a normalization artifact that the researcher can address. Under Assumptions 1 and 2, $\texttt{Var}(x^*) = 1$, so β is the marginal effect of a one standard deviation increase in x^*. The variance of the PCA index equals one only in the formative model with standardized and orthogonal manifest variables.[17] To make marginal effects more comparable, we recommend focusing on the effect of a one standard deviation change in x^{PCA}, equal to $\widehat{\beta}\sqrt{\texttt{Var}(^{\text{PCA}})}$.

The second category is *irrecoverable* and persists regardless of rescaling. The variance-maximization objective of PCA is fundamentally misaligned with the reliability-maximization objective relevant for

[17] Since manifest variables are necessarily correlated under the reflective model, the variance of the PCA index is never guaranteed to be one.

regression: PCA selects the *direction* in manifest-variable space along which total variance is largest, not the direction along which signal variance is largest relative to noise. This directional misalignment cannot be undone by standardizing the resulting index. It manifests in three ways. First, the measurement error induced by PCA is generically nonclassical, so standard intuitions about attenuation bias do not apply. Second, the direction and magnitude of bias depend on other covariates in the model through residual confounding. Third, OLS has different properties depending on whether the measurement model conforms to the reflective or formative model; sign reversal for $\widehat{\beta}$ is possible only in the latter.

In sum, researchers must be much more cautious—or even avoid—using PCA to proxy for a latent variable in a linear regression framework for causal inference. This is true even if the proxy is not of interest. Moreover, researchers must think carefully about the underlying measurement model in addition to the economic model to understand their estimator's properties.

4 Alternative Approaches for Index Creation

Optimal index weights satisfying Proposition 4 are infeasible. However, given PCA's limitations, alternative indices may offer improvements in a regression context. We consider several such alternatives; formal analysis is relegated to the appendix.

4.1 Unit Weights

A simple alternative is the equally-weighted average of the non-standardized z_j, also known as unit weights (e.g., Bobko et al. 2007). This is equivalent to a count index—the sum of the z_j's scaled by J (Filmer and Scott 2012). The index, denoted $x^{\overline{z}}$, sets $\lambda_j = 1$ or $1/J$ for all j.[18] Unit weights are less susceptible to outliers or high-variance manifest variables because the weights are chosen without reference to the data (Bobko et al. 2007).[19] In psychology, Rönkkö et al. (2015, p. 77) refer to this as "the most common approach for constructing composite variables for use in OLS regression analysis" and conclude that "in the event that composite-based approximations to latent variable models are actually needed, there is very little reason to use anything else than unit weighted scales" (p. 82). It is also frequently used in health research where the index is a count of, say, reported health conditions, and in economics where an index of consumption or assets is the sum of several categories.

The properties of OLS using a unit weight index are in Appendix E.1. In the reflective model, $\widehat{\beta}$ is attenuated, converging to β multiplied by the conditional reliability ratio. In the formative model, both attenuation and expansion bias are possible as $\widehat{\beta}$ converges to β times the slope coefficient from the regression of x^* on x. In both cases, the bias of $\widehat{\gamma}$ is in unknown direction and sign reversal is possible.

[18] Two well-known applications include Solon (1992), who uses income averaged over increasing numbers of years to proxy for permanent income, and Ashenfelter and Krueger (1994), who considers averaging education reports from respondents and their twins. Equal weight indices are also common in health, e.g., the frailty index (Rockwood et al. 2017) and food security measurement (Bickel et al. 2000).

[19] The weights are fixed *ex ante*, so a leverage point in one variable cannot distort the weighting scheme, unlike PCA. Extreme z_j values still affect the index linearly; the robustness claim pertains only to weight selection.

4.2 Mean z-Score

A popular alternative to PCA is the mean z-score index, denoted $x^{\bar{\bar{z}}}$ (Kling et al. 2007).[20] This entails applying unit weights to the standardized z_j (Bobko et al. 2007):

$$x^{\bar{\bar{z}}} = \frac{1}{J}\sum\nolimits_j \left(\frac{z_j - \bar{z}_j}{\sigma_{z_j}}\right) = \sum\nolimits_j \left(\frac{1}{J\sigma_{z_j}}\right) z_j - \sum\nolimits_j \left(\frac{1}{J\sigma_{z_j}}\right) \bar{z}_j = \Lambda + \sum\nolimits_j v_j z_j, \tag{16}$$

where $v_j = 1/(J\sigma_{z_j})$ and $\Lambda = -\sum_j v_j \bar{z}_j$. The mean z-score index is a linear combination with weights decreasing in σ_{z_j} and not constrained to sum to one.[21] The properties of OLS using this index are provided in Appendix E.2. The OLS estimates of β and γ are not necessarily attenuated under either measurement model. Sign reversal is ruled out only for β in the reflective model.

4.3 Partial Least Squares

Partial least squares (PLS) is a *supervised* dimensionality reduction method. Rare in economics but well established elsewhere, PLS was developed by Wold in the 1960s (see, e.g., Wold 1982) and is widely used in chemistry, psychometrics, and life sciences.[22] Like PCA, PLS constructs orthogonal linear combinations of manifest variables. Unlike PCA, it is supervised: y factors into the weight selection. The first PLS component $x^{\text{PLS}} = \sum_j v_j z_j$ is obtained by selecting the unit-norm vector $\mathbf{v}$ that maximizes $\mathtt{Cov}(x^{\text{PLS}}, y)$. Subsequent components are derived by residualizing y and each z_j with respect to preceding components and repeating the covariance-maximization step on residuals. This yields indices that jointly maximize explained covariance with y, arguably better for prediction (e.g., Geladi and Kowalski 1986).

Formally, the first PLS component solves

$$\mathbf{v}^{\text{PLS}} = \arg\max \left[\mathtt{Cov}\left(\mathbf{v}'\mathbf{z}, y\right)\right]^2 \tag{17}$$

subject to a normalization constraint. Two options exist: unit norm ($\|\mathbf{v}\| = 1$) or unit variance ($\mathtt{Var}(x^{\text{PLS}}) = 1$). The choice matters. Under unit-norm normalization, weights do not depend on covariances between manifest variables:

$$v_j^{\text{PLS}} = \frac{\mathtt{Cov}\left(z_j, y\right)}{\sqrt{\sum_k \mathtt{Cov}\left(z_k, y\right)^2}}, \qquad j = 1, \ldots, J. \tag{18}$$

Under unit-variance normalization, the weights incorporate these covariances:

$$v_j^{\text{PLS}} = \frac{\sum_k \left(\mathbf{\Sigma}_{\mathbf{zz}}^{-1}\right)_{jk} \mathtt{Cov}(y, z_k)}{\sqrt{\sum_i \sum_k \mathtt{Cov}(y, z_i) \left(\mathbf{\Sigma}_{\mathbf{zz}}^{-1}\right)_{ik} \mathtt{Cov}(y, z_k)}}, \qquad j = 1, \ldots, J. \tag{19}$$

[20] A search on Google Scholar for "mean z-score" AND "index" AND "regression" produces more than 6,800 hits. Accessed 8 December 2025.

[21] Karagiannis and Paleologou (2025) use an equally-weighted average after percentile normalization: $(z_j - \min\{z_j\})/(\max\{z_j\} - \min\{z_j\})$. We do not consider this variant here.

[22] A search on Google Scholar for "partial least squares" AND "index" AND "regression" produces more than 168,000 hits; 19,100 since 2024. Accessed 8 December 2025.

Standard PLS software typically implements the unit-norm solution $\mathbf{v}^{\text{PLS}} \propto \mathbf{c}$. We analyze both normalizations because the unit-variance version arises naturally when the researcher constructs a standardized index ($\mathtt{Var}(x^{\text{PLS}}) = 1$) before entering it in a regression, as is common in applied work. If PLS is applied to standardized manifest variables, the formulas are unchanged except covariances are between y and the standardized variables, and $\mathbf{\Sigma_{zz}}$ becomes the correlation matrix.

Under unit-norm normalization, each PLS weight is proportional to the covariance of the corresponding manifest variable with the outcome. Under unit-variance normalization, each weight is a linear combination of all manifest-variable covariances with the outcome, filtered through $\mathbf{\Sigma}_{\mathbf{zz}}^{-1}$; the weight on z_j therefore depends on the entire covariance structure, not solely on $\mathtt{Cov}(z_j, y)$. Under neither normalization are the weights or their sum constrained to be positive or equal to any specific value. In the reflective model with unit-norm normalization, weights are identical across manifest variables under nondifferential measurement error ($\Pr(z_j \mid x^*, y) = \Pr(z_j \mid x^*)$), since $\mathtt{Cov}(z_j, y) = \mathtt{Cov}(x^*, y)$ for all j. PLS then collapses to an equal weights index up to scale, differing only in that PLS enforces a unit-norm constraint whereas equal weights imposes unit-sum. Under unit-variance normalization with classical measurement error for all j, the weights are proportional to $\mathbf{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\iota}$, where $\mathbf{\Sigma_{uu}}$ is the diagonal covariance matrix of measurement errors and $\boldsymbol{\iota}$ is a vector of ones: the j-th weight is proportional to $1/\sigma_{u_j}^2$, so PLS assigns larger weight to more precisely measured indicators.[23]

The OLS properties using a PLS index are provided in Appendix F. The estimates of β and γ are not necessarily attenuated under either measurement model regardless of standardization. Sign reversal cannot be ruled out for either parameter.

4.4 Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a third dimensionality reduction approach. Dijkstra (2010) contends that factor models are the most commonly used method in the social sciences. EFA identifies the variance shared among manifest variables and attributes it to a parsimonious set of common factors. Unlike PCA or PLS, there is no single closed-form weight vector—different extraction methods (e.g., principal axis factoring, principal components factoring, maximum likelihood) yield different weights. Because principal axis factoring identifies shared variance rather than maximizing total variance as in PCA, it may be better suited for index creation, particularly in the reflective model. However, while PCA performs poorly when measurement error variances are large, EFA performs poorly when measurement errors are highly correlated. In contrast to PLS, EFA does not incorporate the outcome into factor extraction. We focus only on the first factor.

The OLS properties using an EFA index based on principal axis factoring with a regression approach to factor loadings are provided in Appendix G. Because EFA relies on the correlation matrix, we restrict attention to the standardized case. As with PLS, the estimates of β and γ are not necessarily attenuated under either measurement model. Sign reversal cannot be ruled out for either parameter.

[23] By the Sherman–Morrison formula, $\mathbf{\Sigma}_{\mathbf{zz}}^{-1}\boldsymbol{\iota} \propto \mathbf{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\iota}$ when $\mathbf{\Sigma_{zz}} = \mathtt{Var}(x^*)\boldsymbol{\iota\iota}' + \mathbf{\Sigma_{uu}}$. Weights are identical across indicators only when measurement error variances are equal.

4.5 Discussion

Several simplifications were imposed to keep the analysis tractable. First, errors u in the reflective model are assumed independent. Second, manifest variables are assumed to be unbiased reflections of x^* subject to classical measurement error. Third, the numbers of available (J) and true ($\mathcal{J}$) manifest variables are fixed. Some assumptions will be relaxed in Section 6.

Despite these simplifications, several key insights emerge. First, replacing a latent construct with a proxy index generally biases the estimate of β; coefficients on other regressors are nearly always biased as well. Second, proxy error is generally nonclassical, so $\widehat{\beta}$ and $\widehat{\gamma}$ are not guaranteed to be attenuated. Third, bias direction and magnitude depend on the structural relationship between manifest variables and latent construct—researchers must determine whether the reflective or formative model characterizes their structural model. Finally, some indices under some measurement models allow sign reversal for $\widehat{\beta}$. Sign reversal for $\widehat{\gamma}$ is always possible. These findings underscore the importance of thinking carefully about the measurement model alongside the economic model.

Prior to the simulations and applications, we discuss alternatives that avoid constructing a proxy index altogether.

5 Alternatives to Index Creation

5.1 Linear Regression (Lubotsky–Wittenberg Method)

The first approach, developed in Lubotsky and Wittenberg (2006) and extended in Bollinger and Minier (2015), applies to the reflective model with non-standardized manifest variables and measurement errors u_j that may or may not be correlated. As Lubotsky and Wittenberg (2006) show, OLS estimation of

$$y = \alpha + \sum\nolimits_j \beta_j z_j + \gamma w + \varepsilon \tag{20}$$

produces an estimate $\widehat{\beta}^{\text{LW}} := \sum_j \widehat{\beta}_j$ identical to the OLS estimate from the infeasible regression

$$y = \alpha + \beta \widetilde{x}^o + \gamma w + \varepsilon, \tag{21}$$

where $\widetilde{x}^o := \sum_j \widetilde{\lambda}^o z_j$ is the optimal linear index in terms of minimizing the bias of $\widehat{\beta}$ *subject to the unit-sum constraint* under the reflective model. This does not require the researcher to know the measurement error covariance matrix and does not impose $\widetilde{\lambda}^o \in [0, 1]$. Bollinger and Minier (2015) show this procedure also minimizes the bias of $\widehat{\gamma}$. By including manifest variables separately, OLS automatically finds the best linear combination subject to the unit-sum constraint and estimates β and γ accordingly. Note, however, that $\widetilde{x}^o$ differs from the unconstrained optimal index x^o in Proposition 4 and Corollary 1. Thus, OLS does not produce a consistent estimate of β.

Nonetheless, this remarkable result suggests researchers may forego index creation: run a regression with all manifest variables and report the sum of their coefficients as the estimated effect of the latent factor. The plims are derived in Lubotsky and Wittenberg (2006) and Appendix H. However, this has seemingly gone unnoticed despite Lubotsky and Wittenberg (2006, p. 549) warning that creating summary measures such as PCA is "generally ad hoc and hardly ever optimal."[24] It is important to remember this applies only in the reflective model with classical (possibly correlated) measurement error—further evidence that researchers must think critically about the measurement model.

In the formative model, the bias has two components and $\widehat{\beta}^{\texttt{LW}}$ can be attenuated or not. The first component is proportional to the sum of true weights on observed manifest variables. The second reflects adjustment for covariates w and any omitted manifest variables. If $w = 0$ or w and $\mathbf{z}$ are orthogonal, the plim simplifies to $\beta \sum_j \lambda_j^*$.

5.2 Nonlinear Regression (Yang–Jia–Li Method)

A straightforward extension follows from Yang et al. (2023), who consider aggregating high-frequency data to lower frequency. Here, we substitute for $\widetilde{x}^o$ in Equation (21) and impose the unit-sum constraint:

$$y = \alpha + \beta \sum_{j \neq J} \widetilde{\lambda}_j^o z_j + \beta \left(1 - \sum_{j \neq J} \widetilde{\lambda}_j^o\right) z_J + \gamma w + \varepsilon. \tag{22}$$

While nonlinear in $\{\alpha, \beta, \widetilde{\boldsymbol{\lambda}}^o, \gamma\}$, the parameters are identified under usual conditions and can be estimated via Generalized Method of Moments (GMM), Nonlinear Least Squares (NLS), or maximum likelihood. The parameter estimates are identical to the Lubotsky and Wittenberg (2006) approach. However, because the weights $\widetilde{\boldsymbol{\lambda}}^o$ are estimated jointly with α, β, and γ, three possibilities emerge: (i) the optimal index can be generated after-the-fact, (ii) the weights may be constrained to the unit interval, and (iii) the optimal weights can be compared to PCA or alternative weights.

5.3 Instrumental Variables

It is well known that β can be consistently estimated by IV in the reflective model using one or more manifest variables to instrument for another if measurement errors are uncorrelated across variables; combining multiple IV estimates yields more efficient estimates (e.g., Andersson and Möen 2016; Gillen et al. 2019).[25] However, this is difficult to justify in practice, particularly with bounded manifest variables (Black et al. 2000).[26] If independence is violated, the probability limit is

$$\texttt{plim}\, \widehat{\beta}^{\texttt{IV}} = \beta \left[\frac{1}{1 + \texttt{Cov}\left(u_j, u_k\right)}\right] \tag{23}$$

in the regression model with no other covariates, where we normalize $\texttt{Var}(x^*) = 1$.[27] The IV estimate suffers from attenuation bias if $\texttt{Cov}(u_j, u_k) > 0$, expansion bias otherwise. However, even when

[24] Lubotsky and Wittenberg (2006) has only 222 citations on Google Scholar as of 12 December 2025.

[25] Black and Smith (2006) show IV inconsistency even with independent measurement errors if manifest variables are on different scales than x^*. They develop a method to overcome this by first estimating each manifest variable's scale under independent errors.

[26] In a canonical example, Ashenfelter and Krueger (1994) consider using a twin's report of their sibling's education as an instrument for the sibling's self-reported education.

[27] In multiple regression, Frisch–Waugh–Lovell implies the plim takes the more general form $\beta\,\texttt{Var}(x^* \mid \mathbf{w})/[\texttt{Var}(x^* \mid \mathbf{w}) + \texttt{Cov}(u_j, u_k)]$, since $\texttt{Var}(x^* \mid \mathbf{w}) \leq 1$ replaces the unit normalization in Equation (23). Under Assumption 1, $\texttt{Cov}(u_j, u_k \mid \mathbf{w}) = \texttt{Cov}(u_j, u_k)$.

consistent, IV does not necessarily dominate the Lubotsky and Wittenberg (2006) approach in mean squared error, particularly with weak instruments or in the presence of high leverage observations, clustering, and heteroskedasticity (e.g., Lubotsky and Wittenberg 2006; Young 2022).

In the formative model, all observed manifest variables belong in the regression through x^*. Setting aside one or more manifest variables violates the exclusion restriction. In principle, IV remains valid if a variable q exists that is correlated with observed manifest variables $(z_1, ..., z_J)$ and uncorrelated with unobserved ones $(z_{J+1}, ..., z_{\mathcal{J}})$. Finding such a variable in practice is difficult.

An IV approach that can yield consistent estimates is the hybrid approach in Blundell et al. (2023), based on the Multiple Indicators Multiple Causes (MIMIC) framework (Jöreskog and Goldberger 1975). This approach posits the existence of manifest variables decomposable into reflective and formative indicators (see Figure 2). An index derived from the formative indicators may then be a valid instrument for a linear index derived from the reflective indicators.

Formally, let the reflective indicators be $z_j = x^* + u_j$, where u_j is classical measurement error uncorrelated with x^* and w as before. The latent x^* is determined from the formative model:

$$x^* = \sum\nolimits_{\ell=1}^{\mathcal{L}} \lambda_\ell^* q_\ell. \tag{24}$$

Based on the $J \leq \mathcal{J}$ and $L \leq \mathcal{L}$ observed variables, two linear indices are created:

$$x = \sum\nolimits_{j=1}^{J} v_j z_j \qquad \text{and} \qquad r = \sum\nolimits_{\ell=1}^{L} \lambda_\ell q_\ell. \tag{25}$$

Let $S_v := \sum_{j=1}^{J} v_j$. Replacing x^* with x and using r as the instrument yields estimates converging to

$$\texttt{plim}\, \widehat{\beta}^{\texttt{IV}} = \frac{\beta}{S_v} \tag{26}$$

$$\texttt{plim}\, \widehat{\gamma}^{\texttt{IV}} = \gamma \tag{27}$$

as shown in Appendix I. Thus, the IV estimate of β is consistent up to known scale; $S_v \cdot \widehat{\beta}^{\texttt{IV}}$ is consistent for β. The IV estimate of γ is also consistent. Because the instrument is constructed from formative indicators, unobserved q_ℓ ($\ell = L + 1, ..., \mathcal{L}$) cause no bias.

The difficulty lies in classifying variables as reflective versus formative. In the context of individual health, Blundell et al. (2023) argues that subjective measures are reflective while objective measures are formative. Despite this difficulty, the hybrid approach is a viable tool in the right context.

Without the hybrid approach, using manifest variables to instrument for each other is inconsistent when measurement errors are correlated. However, several lesser-known strategies are available to researchers. First, recent advances allow invalid ("imperfect") instruments: Conley et al. (2012), Kippersluis and Rietveld (2018), and Chalak and Kim (2024) allow the instrument to directly affect y, while Nevo and Rosen (2012) allows correlation with the structural error. Second, Ashley and Parmeter (2015), Kiviet (2016), Kiviet (2020), Kripfganz and Kiviet (2021), and Kiviet (2023) assess what can be learned from OLS in the presence of endogeneity without valid instruments. Third, one might exploit higher moments for identification (Klein and Vella 2010; Lewbel 2012; Lewbel et al. 2024). Finally, Kim and Wilhelm (2024) develop a more powerful t-test for $\beta = 0$ with two manifest variables having correlated measurement errors, searching over all linear combinations for the index and instrument. We do not examine these approaches here as there nothing special about their application in the current context.

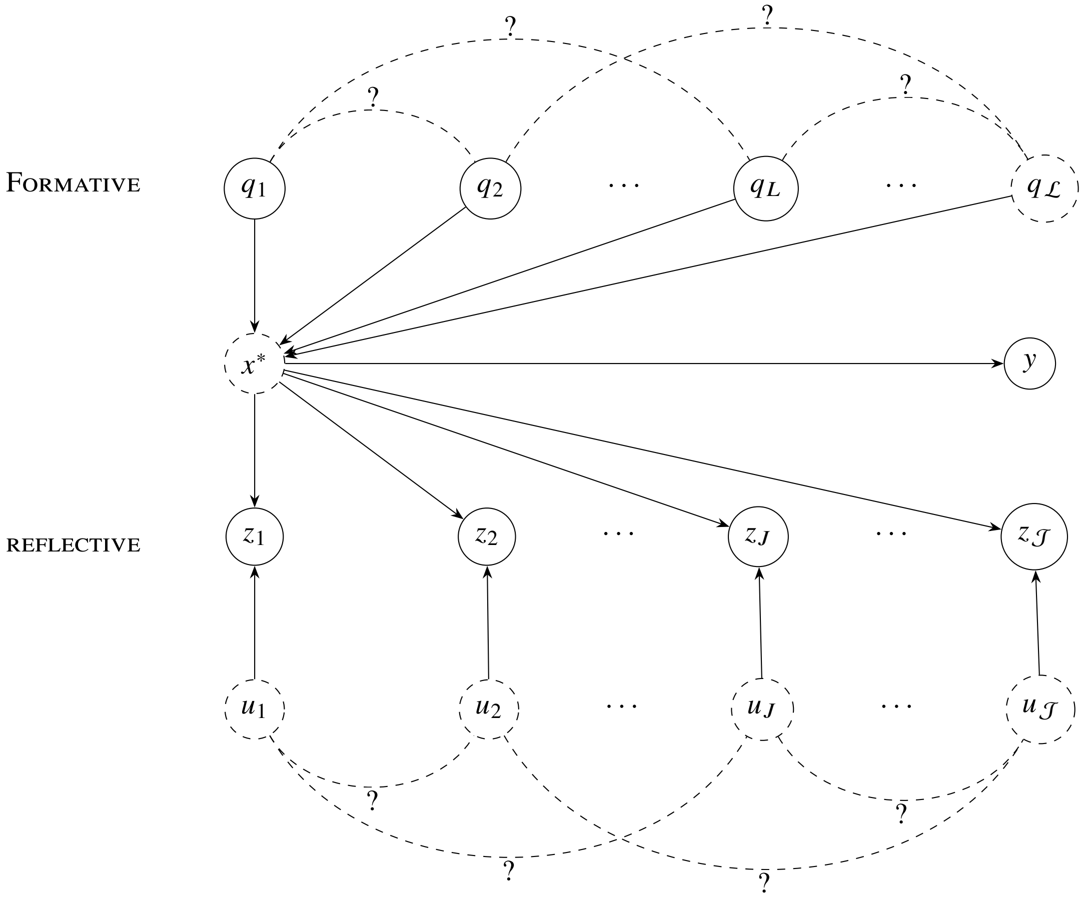


FIGURE 2
HYBRID MEASUREMENT MODEL

NOTES.— Directed Acyclic Graphs (DAG) depicting the relationship between the reflective manifest variables, z_j, $j = 1, ..., \mathcal{J}$, the formative manifest variables, q_ℓ, $\ell = 1, ..., \mathcal{L}$, and x^*. As in Figure 1, dashed circles denote unobserved variables; only the first J reflective and first L formative indicators are observed. Other observed determinants of y (w in Equation (1)) are omitted for simplicity.

5.4 Structural Approaches

Rather than constructing a composite index ex ante, structural approaches model the latent construct explicitly within a joint estimation framework. Blau and Gilleskie (2001) pioneered this strategy in health economics by treating each observed health variable as an imperfect measure of an underlying latent health stock. Their framework captures correlations among health measures and labor-market outcomes through unobserved heterogeneity entering multiple equations simultaneously, allowing joint estimation to discipline how each indicator contributes to inference about latent health.

Our approach differs fundamentally. Whereas Blau and Gilleskie (2001) embed health measurement inside a fully specified structural system and rely on joint estimation to recover latent health's role, we operate in a reduced-form regression environment treating the latent construct as a nuisance parameter. The focus is not on modeling health dynamics or behavioral responses but on understanding how alternative proxies affect coefficient estimation, inference, and conclusions. We do so not because one is more important than the other, but because the vast majority of empirical research involving latent constructs relies on such reduced-form regressions.

Thus, our contribution complements the joint-estimation tradition. Blau and Gilleskie (2001) demonstrate that using a single observed health measure can severely understate health's effect when relevant dimensions are omitted; we show that commonly used proxy indices can generate equally consequential distortions when weights are chosen mechanically. In empirical contexts—such as our political-economy application—where fully structural modeling is infeasible or undesirable, our discussion provides a principled way to aggregate information while making the measurement–identification tradeoff explicit.

Structural Equation Modeling (SEM) provides a general econometric formalization of joint-estimation logic. Developed from early work on path analysis (Wright 1921) and general latent variable models (Jöreskog 1970), SEM estimates measurement and economic models simultaneously. Unlike approaches treating the composite x as an observed regressor subject to nonclassical measurement error, SEM explicitly incorporates x^* and separates its variance from measurement error (Bollen 1989). This yields consistent estimates of β and γ by modeling the error structure directly, typically via maximum likelihood.

Identification in SEM requires $J \geq 3$ manifest variables (otherwise additional restrictions are necessary) and assumes measurement error orthogonality to both the structural error and model covariates, plus independence across indicators. Recent applications embed SEM within complex dynamic structural models (e.g., Cunha et al. 2010; Agostinelli and Wiswall 2025).[28]

6 Monte Carlo Study

The simulations assess the magnitudes and directions of biases under different index approaches in settings likely encountered in practice. We examine performance for estimating the effects of both x^* and w on y, under the assumptions of the prior sections and under more general conditions.

6.1 Simulation Design

6.1.1 Reflective-Indicators Model

For the reflective model, we assess four experimental designs nested in the following DGP:

$$y_i = \beta x_i^* + \gamma w_i + \varepsilon_i, \quad i = 1, 2, \ldots, N; \tag{28}$$

$$z_{ji} = \omega_j x_i^* + u_{ji}, \tag{29}$$

similar to Andersson and Möen (2016). In all experiments, $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We vary:

- Sample size: $N \in \{500, 2000\}$

[28] We do not include SEM in our Monte Carlo simulations for two reasons. First, proxy variable methods dominate empirical practice across disciplines. Second, SEM models frequently encounter convergence difficulties in simulation settings with multiple DGPs.

- Parameter values: $\beta \in \{0, 0.25\}$, $\gamma \in \{0, 0.25\}$
- Covariates: $x^*, w \overset{\text{iid}}{\sim} \mathcal{N}_2(\Upsilon, \boldsymbol{\Sigma}_{\mathbf{x^*w}})$ where $\Upsilon = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\boldsymbol{\Sigma}_{\mathbf{x^*w}} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$
- Scaling of manifest variables: $\{\omega_1, \omega_2, \omega_3\} = \{1, 1, 1\}, \{1, 1.5, 0.5\}, \{1, 5, 0.5\}, \{1, 10, 0.5\}$
- $u_1, ..., u_J \overset{\text{iid}}{\sim} \mathcal{N}_J(0, \boldsymbol{\Sigma}_{\mathbf{uu}})$
- Number of manifest variables: $J \in \{3, 5, 10, 25, 50\}$

where $\mathcal{N}_k$ denotes a k-dimensional multivariate normal. The smaller N aligns with our application; the larger aligns with studies creating individual- or household-level indices. Varying β and γ explores the role of additional exogenous covariates correlated with x^* and allows examination of Type I and Type II errors. We vary scaling as many studies combine manifest variables measured in different units. Finally, varying J and $\boldsymbol{\Sigma}_{\mathbf{uu}}$ explores consequences of varying the number and noisiness of manifest variables.

For each design, we perform 500 simulations and report: (1) mean bias, (2) root mean squared error (RMSE), (3) empirical coverage probability of nominal 95% confidence intervals, and (4) empirical rejection rate of the null that the coefficient equals zero using a two-sided 5% test. We report these for β, η, and γ, where η is the marginal effect of an empirical one standard deviation increase in the index, permitting better assessment of indices that are not unit variance.

We compare several estimators: (1) OLS using latent x^* as a benchmark, (2) OLS using a PCA index with and without standardization, (3) OLS using an equally weighted average without standardization (Unit Weight Index) and with standardization (Mean z-score Index), (4) OLS using a PLS index with and without standardization under unit variance normalization, (5) OLS using an EFA index with standardization, (6) OLS including all manifest variables as covariates and computing the equally-weighted or weighted sum of estimates (LW and LW Weighted; Lubotsky and Wittenberg (2006)),[29] (7) GMM including all manifest variables and estimating optimal index weights—subject to the unit-sum constraint, with and without restricting individual weights to the unit interval—along with slope coefficients (YJL and YJL Weights $\in [0, 1]$; Yang et al. (2023)), and (8) IV via two-stage least squares using each combination of $J - 1$ manifest variables to instrument for the remaining one.[30]

The four experimental designs are:

- Scenario 1: $J = 3$, $\gamma = 0$, $\omega_j = 1 \; \forall j$, $N \in \{500, 2000\}$, and

$$\boldsymbol{\Sigma}_{\mathbf{uu}} = \begin{bmatrix} 0.5 & 0.0 & 0.0 \\ 0.0 & 5.0 & 0.0 \\ 0.0 & 0.0 & 10.0 \end{bmatrix}.$$

Each manifest variable z_j is a noisy indicator of x^* with different noise levels. These values correspond to reliability ratios $\rho_j = \texttt{Var}(x^*)/\texttt{Var}(z_j)$ ranging from roughly 0.09 to 0.67. Within

[29] Lubotsky and Wittenberg (2006) propose a weighted sum when manifest variables are not unbiased reflections of x^*. The weight on z_1 is normalized to one; weights on z_j equal $\texttt{Cov}(z_j, y)/\texttt{Cov}(z_1, y)$, $j = 2, ..., J$. This has been criticized since weights vary with y (Starr 2019).

[30] Simulations performed in `Stata 19` using `pca`, `pls`, `factor`, `regress`, `gmm`, and `ivregress 2sls`.

each design, z_1 is relatively high quality, z_2 is medium quality, and z_3 is extremely noisy—reflecting the common situation where researchers have manifest variables of varying reliability.

- Scenario 2: Identical to Scenario 1 except $\gamma = 0.25$. An additional covariate correlated with the latent construct is included, as typical in regression applications. For example, Radatz and Baten (2025) explore an index of inequality (Gini coefficients for income, height, and land) on civil conflict conditional on institutional quality and ethnic fractionalization. We repeat this scenario with $\gamma = 0$ (retaining the covariance matrix for x^* and w) to assess size distortion.

- Scenario 3: Identical to Scenario 2 except $\omega_1 = 1$, $\omega_2 \in \{1.5, 5, 10\}$, and $\omega_3 = 0.5$. We consider only $N = 500$. The second and third manifest variables are on different scales than x^*. For example, Montero and Yang (2022) form a municipality-level development index where manifest variables are measured in people (population), currency (income), years (education), and percentages (employment rate). Maccini and Yang (2009) construct an asset index combining currency (asset values) and binary indicators (own a television). Black and Smith (2006) consider college quality proxies including a ratio (faculty per student), percentages (rejection and retention rates), points (SAT scores), and currency (faculty salaries).

- Scenario 4: Same as Scenario 2 except $J \in \{3, 5, 10, 15, 25, 50\}$ and

$$z_{ji} = x^* + \frac{j}{5} u_{ji}, \quad j = 1, ..., J,$$

where $u_1, ..., u_J \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We consider only select estimators and $N = 500$. Manifest variables are increasingly noisy reflections of x^*, allowing investigation of how additional low-signal manifest variables influence estimators. For example, Blattman et al. (2013) construct an income/consumption index using 70 manifest variables, Montero and Yang (2022) use 35, Levai and Turati (2025) use 36 labor laws for worker protection, Garmaise (2009) uses 12 for non-compete enforceability, and Woessmann (2025) discusses the vast number available for human capital. Stoetzer et al. (2025) find efficiency improves as reliable manifest variables increase; however, Black and Smith (2006) obtain estimates closer to zero with five versus two manifest variables.

6.1.2 Formative-Indicators Model

For the formative model:

- Scenario 5: Identical to Scenario 1 except $\gamma = 0.25$ and x^*, w, and the z's are generated as

$$\begin{aligned} w_i &= u_i + u_{wi} \\ z_{ji} &= u_i + u_{ji}, \; j = 1, 2, \ldots, \mathcal{J} \\ x^* &= \sum_{j=1}^{\mathcal{J}} \lambda_j^* \tilde{z}_j \end{aligned}$$

where $u, u_w, u_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $\lambda_j \overset{\text{iid}}{\sim} \mathcal{U}[0, 1]$, $\lambda_j^* = \lambda_j / \sigma$ with σ chosen to normalize $\mathtt{Var}(x^*) = 1$,

$\tilde{z}_j$ is standardized z_j, $\mathcal{J} \in \{3, 5, 20\}$, and only z_1, z_2, z_3 are observed. We consider only $N = 500$. Here x^* is a weighted average of all $\mathcal{J}$ manifest variables with randomized weights. When $\mathcal{J} = 3$, all are observed; other cases allow increasing numbers of unobserved manifest variables. We repeat with $\gamma = 0$ to assess size distortions.

6.2 Simulation Results

Complete results are in Appendix J. We focus on a few salient findings.

First, researchers want to know whether an effect exists. At minimum, procedures should have sufficient power to reject the null of zero when false and be appropriately sized when true. Table 1 reports power across Scenarios 1, 2, 3, and 5 when β, η, and γ are 0.25—this is $1 - \Pr(\text{Type II error})$, the probability of correctly rejecting the null. The table also summarizes coverage when true values are zero—this is $1 - \Pr(\text{Type I error})$, the probability of correctly failing to reject the null of zero.

The key takeaway is that no single approach is "best" across all DGPs. When the latent covariate matters ($\beta = 0.25$), most index methods exhibit high power (> 89%); the Unit Weight Index and non-standardized PCA are exceptions. While non-index approaches like LW and YJL can achieve high power in specific settings, their minimum power is strikingly low (roughly 12–13%), suggesting they are highly sensitive to the underlying measurement model. When the latent covariate does not matter ($\beta = 0$), most methods are sized appropriately with the exceptions of PLS and LW: Weighted. These approaches yield high false positive rates (> 25%) across all DGPs. When the focus is on the other covariate in the model, all methods except for IV exhibit very high power (> 99%) when the covariate matters ($\gamma = 0.25$). However, all non-IV methods yield high false positive rates (> 40%) under some measurement models; IV produces false positive rates as high as nearly 25% in some cases.

TABLE 1
SUMMARY OF POWER AND COVERAGE RATES ACROSS ALL SCENARIOS

	β				η				γ			
	Power ($\beta = 0.25$)		Coverage ($\beta = 0$)		Power ($\eta = 0.25$)		Coverage ($\eta = 0$)		Power ($\gamma = 0.25$)		Coverage ($\gamma = 0$)	
Method	min	max	min	max	min	max	min	max	min	max	min	max
Panel A. Benchmark												
True x^*	0.968	1.000	0.936	0.960	0.998	1.000	0.936	0.960	0.998	1.000	0.940	0.966
Panel B. Index Approaches												
Unit Weight Index	0.762	1.000	0.938	0.948	0.762	1.000	0.938	0.948	1.000	1.000	0.032	0.950
EFA Index	0.894	1.000	0.934	0.946	0.894	1.000	0.934	0.946	1.000	1.000	0.328	0.952
Mean z-score Index	0.898	1.000	0.932	0.950	0.898	1.000	0.932	0.950	1.000	1.000	0.194	0.948
PCA Index (Non-Std)	0.304	0.998	0.948	0.950	0.304	0.998	0.948	0.950	1.000	1.000	0.000	0.950
PCA Index (Std)	0.896	1.000	0.926	0.950	0.896	1.000	0.926	0.950	1.000	1.000	0.266	0.948
PLS Index (Non-Std)	0.900	1.000	0.744	0.750	0.900	1.000	0.744	0.750	1.000	1.000	0.198	0.954
PLS Index (Std)	0.906	1.000	0.718	0.748	0.906	1.000	0.718	0.748	1.000	1.000	0.490	0.958
Panel C. Non-Index Approaches												
IV ($z_1, z_2 \to z_3$)	0.686	1.000	0.954	0.962	0.688	1.000	0.954	0.962	0.734	1.000	0.758	0.966
IV ($z_1, z_3 \to z_2$)	0.824	1.000	0.948	0.956	0.826	1.000	0.948	0.956	0.828	1.000	0.784	0.958
IV ($z_2, z_3 \to z_1$)	0.562	1.000	0.942	0.948	0.562	1.000	0.942	0.948	0.816	1.000	0.772	0.968
LW	0.124	1.000	0.942	0.948					1.000	1.000	0.574	0.946
LW: Weighted	0.910	1.000	0.694	0.714					1.000	1.000	0.566	0.950
YJL	0.134	1.000	0.950	0.952	0.134	1.000	0.950	0.952	1.000	1.000	0.576	0.958
YJL: Weights $\in [0, 1]$	0.310	1.000	0.912	0.920	0.310	1.000	0.912	0.920	0.992	1.000	0.574	0.956

NOTES.—min and max values computed across Scenarios 1, 2, 3, and 5. Power calculated when true parameter is 0.25 (Scenarios 1, 2, 3, 5 for β and η; Scenarios 2, 3, 5 for γ). Coverage calculated when true parameter is 0 (Scenario 1 for β and η; Scenarios 2, 3, 5 for γ). η not estimated for LW methods.

Focusing on two popular approaches in economics—Mean z-score and standardized PCA—should frighten researchers. While both do a reasonable job in terms of power and coverage related to the latent construct and an excellent job in terms of power for the other covariate in the model, they produce false positive rates for γ as high as 81% and 74%, respectively. Thus, researchers *cannot ignore measurement issues* even when they are not directly related to the object of interest.

Figure 3 plots coverage and power for select estimators under Scenario 4, varying J from 3 to 50. The main insight is that performance of the Unit Weight Index and to a lesser extent the Mean z-score Index deteriorate rapidly as J increases. For remaining approaches, increasing J either does not affect performance or causes modest decline. With the caveat that results are specific to our DGP, there appears to be no gain in coverage or power from using many manifest variables when signal declines with J.

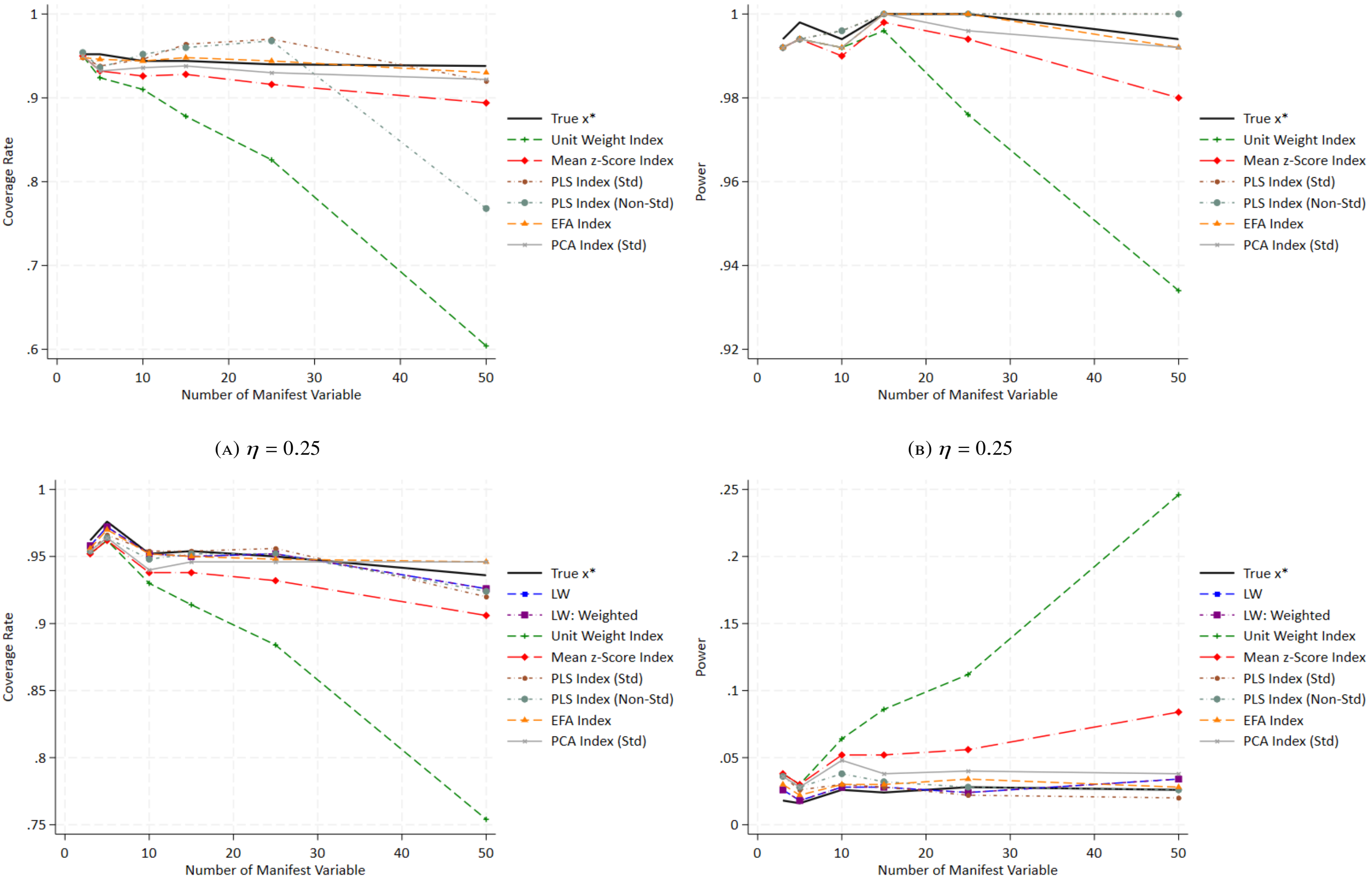
Second, detailed results in Appendix J reveal deeper insights regarding bias and efficiency:

- **Covariate Leakage:** Across all scenarios including γ, a clear pattern emerges: measurement error in x^* "leaks" into $\widehat{\gamma}$. Even when w is perfectly measured and uncorrelated with the errors in the manifest variables, using a proxy index induces bias in $\widehat{\gamma}$. This leakage is most severe in the PCA and Mean z-score approaches, which prioritize capturing variance in z over purging the specific errors that contaminate the structural regression.

- **Reflective Models:** In Scenarios 1 and 2, index methods are dominated by IV in large samples. While indices like PCA and Mean z-score are more efficient (lower RMSE), they suffer from persistent, non-trivial bias that does not vanish as N increases. The bias tends to be toward zero for β (attenuation), but away from zero for γ (leakage). This suggests that researchers using proxy indices are not merely distorting size; they are systematically over-reporting the importance of their control variables by attributing the unmeasured influence of the latent construct to other covariates. While IV achieves near-zero bias and nominal 95% coverage, its estimates are significantly noisier, particularly for the standardized effect η.

- **Formative Models:** The hierarchy flips in Scenario 5. Here, IV is dominated by LW: Weighted and PLS, which exhibit nearly zero bias and superior efficiency. Because the manifest variables cause x^*, the instruments are invalid, leading to poor coverage and higher RMSE. As the ratio of observed to total indicators ($J/\mathcal{J}$) declines, non-IV performance worsens because the signal falls, yet IV paradoxically improves in bias and coverage. This confirms the importance of thinking critically about all aspects of the measurement model; applying reflective model corrections to a formative DGP fundamentally distorts the estimates.

- **The Scaling Trap:** Scenario 3 demonstrates that non-standardized indices (PCA and Equal Weight) are effectively unusable when manifest variables have different scales. In these cases, the magnitude of $\widehat{\eta}$ becomes essentially uninformative, with RMSEs nearly as large as the parameter values. Conversely, PLS and EFA adapt well to scaling differences, matching the performance of standardized indices.



(A) $\eta = 0.25$

(B) $\eta = 0.25$

(C) $\gamma = 0$

(D) $\gamma = 0$

FIGURE 3

SCENARIO 4: COVERAGE AND POWER AS NUMBER OF MANIFEST VARIABLES VARIES ($\eta = 0.25$, $\gamma = 0$)

NOTES.—$N = 500$. Results for select approaches only. Left panels show coverage rates for the true parameter value; right panels show power (rejection rate) when the null is zero.

7 Principles for the Applied Researcher

Composite indices have become commonplace in empirical economics, yet they are often assembled with little attention to the underlying measurement model or econometric implications of the chosen construction. Our results show these choices are far from neutral. In many settings, the index implicitly defines the identifying variation, and mechanically convenient defaults such as PCA can create large distortions. To avoid these pitfalls, researchers must heed the plea in Griliches (1985) to treat measurement issues with the same care typically reserved for research design. With this in mind, we offer the following practical recommendations.

1. *Articulate the measurement model.* Researchers must specify whether the latent construct relates to observed manifest variables via the reflective or formative model. In reflective settings, proxies are noisy manifestations of an underlying latent variable; in formative settings, the construct is an aggregation of its components. Moreover, researchers should consider how the set of observed manifest variables relates to the complete set of manifest variables. As noted in Jarvis et al. (2003) and Edwards and Bagozzi (2000), empirical researchers often devote substantial effort to justifying structural relationships while giving limited attention to measurement relationships. Absent clarity on the measurement model, an index has no coherent interpretation and econometric implications become opaque.

2. *Be transparent about construction choices.* Index construction involves a sequence of decisions—manifest variable selection, standardization, normalization, aggregation, and validation—that affect the resulting proxy and subsequent econometric properties. As emphasized in Greco et al. (2019, p. 63), hidden or *ad hoc* decisions create opportunities for manipulation and hinder replication, while naïve weighting choices may yield indices with no meaningful interpretation. Researchers must document and justify every step.

3. *Assess and report sensitivity.* Because different constructions embody different identifying assumptions, presenting results across a reasonable set of alternatives greatly enhances credibility. As Greco et al. (2019, p. 86) emphasize, robustness checks are “an excellent quality-assurance tool,” even if they cannot substitute for a transparent theoretical framework. Sensitivity analyses should scrutinize all aspects of index construction: manifest variable selection, transformations, normalization, weighting, and aggregation. Highly volatile or poorly measured variables can dominate PCA-style methods and lead to low signal-to-noise ratios; variables with unclear conceptual justification may warrant down-weighting or exclusion. Approaches that down-weight poor proxies, such as PLS under unit-variance normalization, help automate this process, but conceptual scrutiny remains essential. Replication materials must include the manifest variables, weighting formulas, and code sufficient to reproduce indices; do not provide only the constructed index.

4. *Interpret data-driven weights with caution.* Weights from PCA, PLS, EFA, and similar methods are often treated as evidence of relative importance, but data-driven weights “do not necessarily correspond to the actual linkages among the indicators . . . and do not necessarily reflect a sound theoretical framework” (Greco et al. 2019, p. 72). These weights maximize statistical variation or

covariation in the proxy space, not alignment with a behavioral or conceptual model—and certainly not alignment with the econometric objective of consistently recovering a regression coefficient.

5. *Choose estimation methods carefully.* Researchers should select an econometric approach that performs well in their application. Our simulation results can help guide this choice, or researchers can perform new simulations aligned with their specific measurement and economic models (see Figure 4). Sensitivity analyses using alternative estimation strategies should be conducted. As discussed in Section 5, methods beyond those examined here—including the hybrid IV approach of Blundell et al. (2023), imperfect instrument methods, and structural equation modeling—should also be part of the researcher's toolkit.

6. *Consider defining the problem away.* Researchers can always define away any measurement error problem by shifting the focus from x^* to x, where the latter is free of measurement error by construction. To do so requires that (i) researchers not interpret the marginal effect of x as if it were the marginal effect of x^*, and (ii) researchers justify the assumptions required for consistent estimation of their model when the regressor is x instead of x^*.

These principles elevate index construction from a mechanical pre-processing step to an integral component of research design. Composite indices can be powerful tools, but only when built with conceptual clarity, transparent procedures, and careful attention to their econometric implications.

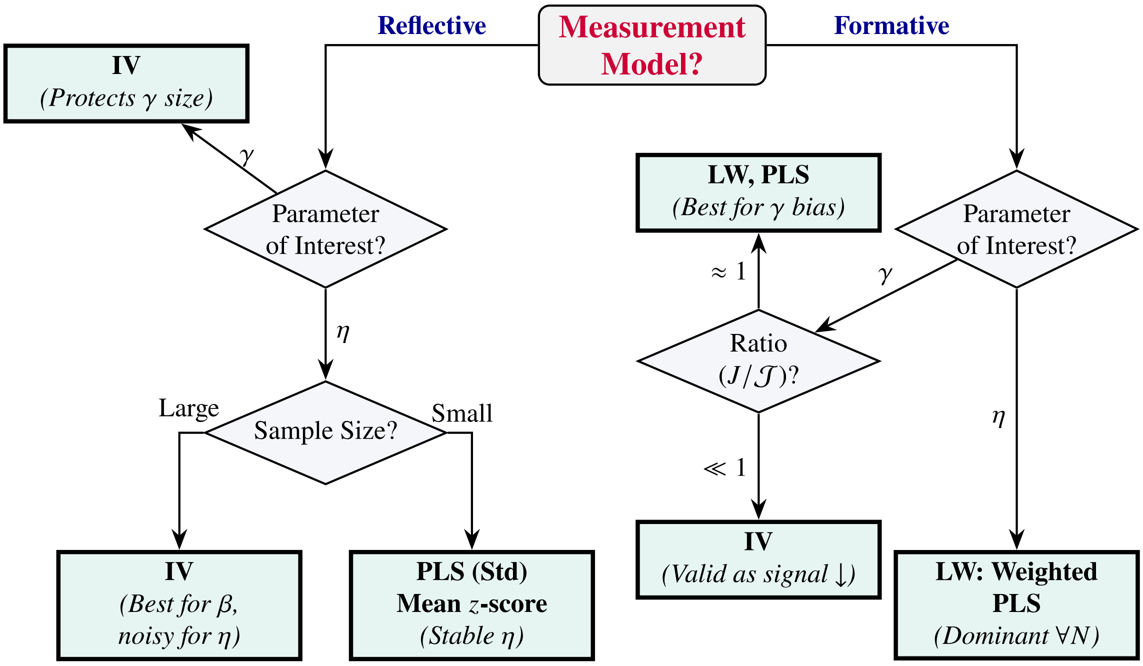


Figure 4
Suggestive Flowchart Based on Simulations

Notes.—IV = Instrumental Variables. PLS = Partial Least Squares. LW = Lubotsky and Wittenberg (2006). η is the marginal effect of a one standard deviation increase in the latent construct. γ is the the coefficient on other covariates in the model. J is the number of observed manifest variables. $\mathcal{J}$ is the total number of manifest variables.

8 Replication: *Overconfidence in Political Behavior*

Ortoleva and Snowberg (2015) study overconfidence in political behavior using two composite indices. A *Media Index* summarizes political information exposure by aggregating survey responses on engagement with traditional media—reading newspapers, watching television news, listening to radio news, and reading blogs—into a scalar proxy for the latent construct "news consumption." Conceptually, this setup corresponds to the formative model as engagement is the (weighted) sum of all news sources accessed. Moreover, because the four media sources capture the dominant traditional news sources, the setup is likely characterized by $J/\mathcal{J}$ close to one. An *Overconfidence Index* is a scalar proxy for the latent construct "belief certainty in excess of actual political knowledge." It is constructed by residualizing self-assessed confidence on accuracy across factual questions about current and expected unemployment and inflation rates, then aggregating the residuals into an index. Conceptually, this setup corresponds to the reflective model as overconfidence determines an individual's beliefs about their knowledge. For both constructs, Ortoleva and Snowberg (2015) employ PCA, using the standardized first principal component as the proxy.[31]

Our goal with this replication is to assess how different index-construction methods may impact an analysis in practice despite the underlying structural model being constant. To proceed, we revisit the authors' analysis using alternative approaches to index construction—Unit Weight, Mean z-scores, PLS, the Lubotsky and Wittenberg (2006) method, and IV—to assess sensitivity of the original findings. Following Figure 4, LW and PLS are our preferred estimators when news consumption is the latent construct given that the manifest variables correspond to the formative model and the ratio $J/\mathcal{J}$ is likely to be close to one. When overconfidence is the latent construct, IV is our preferred approach given that it performs well in the reflective model with manifest variables on the same scale under independent measurement errors. Here, the manifest variables are residualized and thus mean zero and have variances close to one. Moreover, the residualization may absorb much of the correlation in the measurement errors.

Table 2 replicates Table 2 in Ortoleva and Snowberg (2015), varying the *Media Index*. Following Ortoleva and Snowberg (2015) all indices are standardized to unit variance. We study three outcome variables. *Overconfidence* (Panel A) is the index described above.[32] *Ideology* (Panel B) is the respondent's self-placement on the liberal–conservative scale. *Squared Deviation* (Panel C) measures ideological extremity as the squared distance between a respondent's ideology and the mean ideology among co-partisans. Models without (columns 2-6) and with (columns 7-10) additional controls are estimated.

The results using PCA are identical to the original study. Other index construction approaches, along with the LW (non-weighted and weighted) approach, confirm the statistically significant, positive effects of latent news consumption on all three outcomes: latent news consumption is associated with greater overconfidence, conservative ideology, and ideological extremeness. Moreover, consistent with the simulations for the formative model (Scenario 5), the magnitude of the effects are very similar across index approaches. That said, point estimates using the Unit Weight Index are always the smallest, while those using PLS are the largest. LW (non-weighted and weighted) are typically much larger (two to threefold larger) except for non-weighted LW in Panel B. The LW approach also illuminates variation in

[31] Ortoleva and Snowberg (2015) subtract the (weighted) minimum and divide by the (weighted) standard deviation.

[32] Sensitivity to the creation of the index when used as the dependent variable is beyond the scope of the current paper.

sign and statistical significance of the manifest variables. The approach points to a large and statistically significant effect of radio consumption, but a mix of signs and statistical significance across the remaining three sources. For *overconfidence* (Panel A), TV, newspaper, and radio have sizable positive coefficients while the estimated effect of blog use is small and imprecise. For *Ideology* (Panel B), radio has a strong positive effect whereas newspaper and blog have negative effects. For *Squared Deviation* (Panel C), blog, newspaper, and radio each have positive and precisely estimated effects. The effects of TV on *Ideology* and *Squared Deviation* is near zero and statistically indistinguishable from zero.

Table 3 replicates Table 3 in Ortoleva and Snowberg (2015), now varying how the *Overconfidence Index* is operationalized. The outcomes are *Ideology* (Panel A) and *Ideological Extremeness Purged of Economic Controls* (Panel B), defined as the absolute value of the residuals from regressions of ideology on socioeconomic covariates. Models without any controls (columns 2-5), with economic controls (columns 6-9), and with economics controls and controlling for the number of signals (count of media channels from which the respondent received news; columns 10-13) are estimated. The results using PCA are identical to the original study. Other index construction approaches and LW (non-weighted and weighted) provide remarkably similar estimates in terms of both magnitude and statistical significance.[33] Although modest, the PLS estimates are consistently smaller in magnitude. Interestingly, in Panel A the LW approach shows that two manifest variables—those related to inflation, not unemployment—drive the results. In Panel B, the coefficients on the manifest variables are predominantly individually statistically insignificant, despite the sum of the estimates being statistically different from zero at the $p < 0.01$ level.

Finally, while not shown, we also use IV when overconfidence is the latent construct. We estimate 12 IV specifications for each outcome where each of the four manifest variables is used as the proxy and the remaining three variables are used as the instruments, with and without and the economic controls and the economic controls plus the number of signals. When *Ideology* is the outcome, the point estimates vary from 0.213 to 0.319 ($p < 0.01$ and first-stage $F > 120$ in all cases). When *Ideological Extremeness Purged of Economic Controls* is the outcome, the point estimates vary from 0.129 (with all controls) to 0.337 (with no controls; $p < 0.01$ and first-stage $F > 120$ in all cases). Thus, the IV estimates reinforce the other approaches and indicate that latent political overconfidence is associated with more conservative ideology and ideological extremeness. However, they also point to effects that are larger in magnitude depending on the choice of instruments.

Taken together, Tables 2 and 3 yield a common conclusion: the sign of the behavioral relationships emphasized by Ortoleva and Snowberg (2015) and the marginal effects of a one standard deviation change in the indices are quite stable, but forgoing index construction with the LW approach can alter the magnitude. Moreover, reporting estimates on individual manifest variables enhances transparency about channel heterogeneity that composite indices suppress.

[33] We omit the Unit Weight Index for space.

Table 2
Replication: Sensitivity to Creation of Media Index

	PCA	Equal Weight	Mean z-Score	PLS	LW	PCA	Unit Weight	Mean z-Score	PLS	LW
Panel A. Overconfidence										
Media Index	0.205***	0.184***	0.212***	0.223***		0.149***	0.131***	0.148***	0.192***	
	(0.041)	(0.036)	(0.041)	(0.042)		(0.038)	(0.033)	(0.037)	(0.039)	
Blog					0.025					0.053
					(0.066)					(0.071)
TV					0.271***					0.127*
					(0.085)					(0.073)
Newspaper					0.169***					0.073
					(0.055)					(0.054)
Radio					0.254***					0.246***
					(0.067)					(0.062)
LW Estimate					0.719***					0.500***
LW SE					(0.143)					(0.132)
LW Estimate (Weighted)					0.618***					0.424***
LW SE					(0.113)					(0.102)
Controls	N	N	N	N	N	Y	Y	Y	Y	Y
Observations	2927	2927	2927	2927	2927	2927	2927	2927	2927	2927
Panel B. Ideology										
Media Index	0.059**	0.054**	0.059**	0.189***		0.055**	0.048**	0.050**	0.191***	
	(0.023)	(0.021)	(0.024)	(0.031)		(0.022)	(0.020)	(0.023)	(0.027)	
Blog					−0.139***					−0.087*
					(0.049)					(0.051)
TV					0.071					−0.001
					(0.066)					(0.061)
Newspaper					−0.141***					−0.167***
					(0.047)					(0.046)
Radio					0.374***					0.385***
					(0.056)					(0.052)
LW Estimate					0.165*					0.130*
LW SE					(0.086)					(0.075)
LW Estimate (Weighted)					0.454*					0.440*
LW SE					(0.073)					(0.068)
Controls	N	N	N	N	N	Y	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868
Panel C. Squared Deviation										
Media Index	0.288***	0.242***	0.271***	0.301***		0.189***	0.158***	0.177***	0.215***	
	(0.028)	(0.024)	(0.028)	(0.029)		(0.028)	(0.024)	(0.028)	(0.025)	
Blog					0.346***					0.270***
					(0.069)					(0.056)
TV					−0.002					−0.008
					(0.059)					(0.059)
Newspaper					0.263***					0.147***
					(0.060)					(0.051)
Radio					0.313***					0.200***
					(0.056)					(0.044)
LW Estimate					0.920***					0.609***
LW SE					(0.100)					(0.101)
LW Estimate (weighted)					0.849***					0.565***
LW SE					(0.081)					(0.072)
Controls	N	N	N	N	N	Y	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868

Notes.—PCA = Principal Component Analysis. PLS = Partial Least Squares. LW = Lubotsky and Wittenberg (2006). PCA columns replicate Table 2 in Ortoleva and Snowberg (2015). Regressions are weights and standard errors are clustered by age. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 3
Replication: Sensitivity to Creation of Overconfidence Index

	PCA	Mean z-Score	PLS	LW	PCA	Mean z-Score	PLS	LW	PCA	Mean z-Score	PLS	LW
Panel A. Ideology												
Overconfidence Index	0.218***	0.215***	0.197***		0.220***	0.217***	0.198***		0.199***	0.197***	0.180***	
	(0.027)	(0.027)	(0.024)		(0.023)	(0.023)	(0.021)		(0.023)	(0.024)	(0.021)	
Reported Unemp				−0.011				−0.008				0.000
				(0.025)				(0.025)				(0.025)
Reported Inflation				0.098***				0.110***				0.110***
				(0.036)				(0.034)				(0.033)
Expected Unemp				0.017				0.001				−0.001
				(0.037)				(0.033)				(0.032)
Expected Inflation				0.110***				0.112***				0.087**
				(0.037)				(0.036)				(0.034)
LW Estimate				0.215***				0.215***				0.196***
LW SE				(0.029)				(0.025)				(0.025)
LW Estimate (Weighted)				0.214***				0.217***				0.195***
LW SE				(0.024)				(0.022)				(0.021)
Economic Controls	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Number of Signals	N	N	N	N	N	N	N	N	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868
Panel B. Ideological Extremeness Purged of Economic Controls												
Overconfidence Index	0.234***	0.235***	0.208***		0.174***	0.174***	0.154***		0.122***	0.122***	0.108***	
	(0.028)	(0.028)	(0.025)		(0.027)	(0.027)	(0.024)		(0.026)	(0.026)	(0.023)	
Reported Unemp				0.076**				0.038				0.029
				(0.034)				(0.033)				(0.029)
Reported Inflation				0.063				0.065*				0.046
				(0.038)				(0.035)				(0.031)
Expected Unemp				0.068				0.025				0.028
				(0.043)				(0.038)				(0.035)
Expected Inflation				0.042				0.053				0.024
				(0.042)				(0.038)				(0.034)
LW Estimate				0.249***				0.181***				0.128***
LW SE				(0.030)				(0.029)				(0.029)
LW Estimate (Weighted)				0.247***				0.182***				0.128***
LW SE				(0.029)				(0.028)				(0.028)
Economic Controls	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y
Number of Signals	N	N	N	N	N	N	N	N	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868	2868

Notes.—PCA = Principal Component Analysis. PLS = Partial Least Squares. LW = Lubotsky and Wittenberg (2006). PCA columns replicate Table 3 in Ortoleva and Snowberg (2015). Regressions are weighted and standard errors are clustered by age. * $p < .10$, ** $p < .05$, *** $p < .01$.

9 Application: *Health and Voting*

American politics has entered an era of pronounced geographic polarization (McCarty et al. 2019), characterized by a widening partisan divide between urban and rural areas. Yet, despite the politically charged nature of debates over COVID-19, vaccines, drug prices, and health insurance, the correlation between population health and electoral outcomes remains under-explored. While a growing literature documents the influence of "place" in general (Finkelstein et al. 2021) and local politics in particular (Millimet and Whitacre 2025) on health outcomes, less is known about the reciprocal impact of health on elections. Wasfy et al. (2020) provide a notable exception, finding that a composite index of adverse health conditions at the U.S. county level predicted an *increase* in the Republican vote share between the 2016 and 2018 elections.[34] However, their analysis raises a measurement question central to this paper: do different approaches to index construction lead to divergent substantive conclusions regarding the role of health in electoral choice?

To address this question, we estimate the association between community health and changes in the Republican two-party vote share between the 2020 and 2024 presidential elections using a cross-section of 429 Congressional Districts (CDs).[35] The change in Republican two-party vote share is measured in percentage points.[36] Consistent with Blundell et al. (2023), we define the latent construct of interest as "community health" and observe manifest variables—measuring access, socioeconomic risk, behavioral and physiological health, and health outcomes—that can arguably be categorized as either reflective or formative. The observed health measures come from the Congressional District Health Dashboard.[37] We include two controls standard in empirical work on political geography: the share of adults aged 25 and older with a bachelor's degree or higher (BA+) and log median household income. These come from the 2023 release of the American Community Survey (ACS).

Table 4 reports descriptive statistics. The average change in Republican vote share is 2.8 percentage points, with modest dispersion. We use 20 manifest variables. Variables capturing access to health care, socioeconomic risk, and behavioral and physiological risk factors are categorized as formative indicators of community health. Health outcomes and functional status are categorized as reflective indicators of community health. All health variables are measured in percentages with the exception of years of potential life lost due to premature mortality and deaths due to breast cancer, colorectal cancer, and cardiovascular disease. The formative health indicators exhibit notable cross-district variation in both access and socioeconomic risks (e.g., Medicaid enrollment, uninsured, child poverty) and behavioral and physiological risk factors (e.g., obesity, dental care). The reflective health indicators likewise display significant cross-district variation. This combination of modest outcome dispersion and meaningful heterogeneity in component indicators is precisely the setting where index choices can be pivotal: alternative constructions emphasize different margins of variation, thereby changing the effective regressor.

[34] Wasfy et al. (2020) provide a brief review of other studies examining the association between public health and recent elections.

[35] We exclude at-large districts (AK, DE, ND, SD, VT, WY) and non-voting territories (DC, PR, GU, VI, AS, MP).

[36] Obtained from `https://www.the-downballot.com/p/the-downballots-calculations-of-presidential`.

[37] Obtained from `https://www.congressionaldistricthealthdashboard.org/`.

TABLE 4
SUMMARY STATISTICS

Variable	Mean	Std. Dev.
OUTCOME AND CONTROLS		
Change in Republican Vote Share (%)	2.758	2.273
(log) Median Income (2023 USD)	11.281	0.248
Share of 25+ with Bachelor's degree or higher (%)	34.795	11.048
ACCESS AND SOCIOECONOMIC RISK; FORMATIVE INDICATORS		
Medicaid Enrollment (%)	24.990	9.410
Uninsured (%)	10.052	4.952
Children in Poverty (%)	16.029	6.209
Housing with Potential Lead Risk (%)	23.076	10.284
Primary-Care Shortage Area (%)	7.751	11.967
BEHAVIORAL AND PHYSIOLOGICAL RISK FACTORS; FORMATIVE INDICATORS		
No Routine Checkup (% of adults)	25.914	4.322
No Dental Care (% of adults)	37.108	6.241
Obesity (% of adults)	32.826	4.848
Physical Inactivity (% of adults)	24.838	4.560
Smoking (% of adults)	16.018	3.407
Binge Drinking (% of adults)	17.292	1.970
Diabetes (% of adults)	10.834	1.922
HEALTH OUTCOMES AND FUNCTIONAL STATUS; REFLECTIVE INDICATORS		
High Blood Pressure (% of adults)	31.926	4.385
Frequent Mental Distress (% of adults)	15.788	1.709
Frequent Physical Distress (% of adults)	12.014	1.886
Independent Living Difficulty (%)	5.813	1.253
Breast Cancer Deaths (per 100,000)	20.953	2.439
Colorectal Cancer Deaths (per 100,000)	13.973	2.145
Cardio Deaths (per 100,000)	203.150	39.286
Premature Deaths (years of potential life lost per 100,000)	8494.631	2296.069
Number of Congressional Districts	429	

NOTES.—Means and standard deviations reported for all variables. Observations exclude at-large congressional districts (AK, DE, ND, SD, VT, WY) and non-voting territories (DC, PR, GU, VI, AS, MP).
SOURCES.—Health measures from Congressional District Health Dashboard. Election data from Downballot Project. Education and income from ACS 2023 release.

The structural model is given by

$$\Delta\text{Vote}_i = \alpha + \beta H_i^* + \gamma_1 E_i + \gamma_2 M_i + \varepsilon_i, \tag{30}$$

where ΔVote_i is the change in Republican two-party vote share in percentage points (pp) for district i, H_i^* is latent community health, E_i is BA+ share, and M_i is log median income. Estimation proceeds by replacing H_i^* with H_i, a health index constructed using alternative methods and alternative sets of manifest variables. Standard errors are clustered by state.

Guided by the results in Sections 7, we report estimates using a range of index constructions rather than privileging a single approach. PCA is included because of its prevalence in applied work, not because it is theoretically preferred. Alternative constructions—factor methods, PLS, Lubotsky-Wittenberg, and Yang–Jia-Li (GMM)—are included to explore sensitivity. Finally, IV approaches are used given their superior performance when the focus is on other covariates in the model.

Table 5 assesses the range of estimates across different index construction choices. Each panel and column is a distinct model. There are 42 models in total (7 sets of manifest variables times 6 index approaches). Based on the listing of variables in Table 4, Columns (2)–(4) use the formative indicators to form the index, with Column (2) using the access and socioeconomic risk variables, Column (3) using the behavioral and physiological risk factors, and Column (4) using all formative indicators. Columns (5)–(7) use the reflective indicators to form the index, with Column (5) using the first four health outcomes and functional status variables, Column (6) using the last four health outcomes and functional status variables, and Column (7) using all reflective indicators. Finally, Column (8) uses all 20 manifest variables. To facilitate comparisons, Table 5 reports the marginal effects of the health index on a common scale (Meijer et al. 2025); we report $\eta := \widehat{\beta} \times \sigma_H$, where σ_H is the standard deviation of the constructed index. This normalization maps all estimates to the effect of a one-standard-deviation change in the method-specific health proxy, ensuring comparability despite differences in scaling and weighting.

In practice, researchers might choose any one of these models without raising suspicion. For example, Wasfy et al. (2020) form an index using PCA applied to nine manifest variables: (1) physically unhealthy days, (2) mentally unhealthy days, (3) percent food insecure, (4) teen birth rate, (5) age-adjusted mortality rate, (6) violent crime rate (7) average per capita healthcare costs, (8) percent diabetic, and (9) percent overweight or obese. The authors then normalize PC1 to the unit interval. Note, in our view, (1), (2), and (5) are reflective indicators while the remainder are formative indicators.

Several findings stand out. First, and most importantly, the magnitude, statistical significance, and even the *sign* of the estimated parameters varies widely. This underscores that index construction is not an innocuous data pre-processing step, but rather an integral part of research design. Second, the variation is more extreme *across columns* than across *rows*. In particular, estimates can diverge markedly depending on whether the formative indicators or reflective indicators or all indicators are used.

Third, models using indices derived from the formative indicators suggest a positive and statistically significant association between (poor) community health and the change in the two-party Republican vote share. A one standard deviation change in the various indices is associated with a 0.27pp (Panel E, Column (3)) to a 2.3pp (Panel D, Column (4)) increase. Using the formative indices also suggest a positive and statistically significant association between median income and two-party Republican vote share; the association between education and two-party Republican vote share is negative, but often statistically insignificant.

Fourth, models using indices derived from the reflective indicators or the combined set of indicators suggest an association between (poor) community health and the change in the two-party Republican vote share that ranges from *negative* and statistically significant at the $p < 0.01$ level to *positive* and statistically significant at the $p < 0.01$ level. A one standard deviation change in the various indices is associated with a 1.3pp *decline* (Panel F, Column (6)) to a 2.1pp *increase* (Panel D, Column (8)). Using the reflective or combined indices also suggests a negative and statistically significant association between education and two-party Republican vote share; the association between median income and two-party Republican vote share ranges from negative and statistically insignificant to positive and statistically significant.

Table 5

Change in Republican Vote Share and Health Indices: Index-Comparison Estimates (Comparable η Scale)

	Formative Indices			Reflective Indices			Combined Index
	$J = 5$	$J = 7$	$J = 12$	$J = 4a$	$J = 4b$	$J = 8$	$J = 20$
A. Mean z-score							
η	1.646***	0.953***	1.996***	−0.688	−1.000***	−1.156**	0.472
	(0.321)	(0.369)	(0.357)	(0.427)	(0.334)	(0.459)	(0.463)
Share BA+ (%)	−0.048*	−0.037	−0.011	−0.092**	−0.090**	−0.099**	−0.061*
	(0.027)	(0.034)	(0.028)	(0.043)	(0.038)	(0.039)	(0.035)
(log) Median income	5.938***	4.062**	6.689***	0.406	−0.129	−0.800	3.410
	(1.640)	(1.766)	(1.691)	(1.879)	(1.770)	(1.989)	(2.352)
B. PCA							
η	1.766***	0.927**	1.969***	1.969***	−1.004***	−1.153**	0.095
	(0.330)	(0.427)	(0.431)	(0.431)	(0.335)	(0.459)	(0.347)
Share BA+ (%)	−0.040*	−0.050	−0.023	−0.023	−0.090**	−0.099**	−0.072*
	(0.021)	(0.053)	(0.045)	(0.045)	(0.038)	(0.039)	(0.042)
(log) Median income	6.283***	4.541***	7.303***	7.303***	−0.148	−0.808	2.441
	(1.559)	(1.218)	(1.551)	(1.551)	(1.769)	(1.988)	(1.764)
C. EFA							
η	1.994***	1.461***	2.166***	−0.086	−1.046***	−1.132**	0.187
	(0.331)	(0.462)	(0.418)	(0.352)	(0.340)	(0.463)	(0.347)
Share BA+ (%)	−0.054***	−0.035	−0.020	−0.077*	−0.091**	−0.098**	−0.070*
	(0.019)	(0.054)	(0.042)	(0.044)	(0.037)	(0.039)	(0.042)
(log) Median income	7.833***	5.883***	7.839***	1.980	−0.293	−0.827	2.696
	(1.552)	(1.215)	(1.337)	(1.724)	(1.751)	(2.019)	(1.760)
D. PLS							
η	2.050***	1.942***	2.318***	1.061***	−1.160***	−1.191***	2.118***
	(0.236)	(0.278)	(0.272)	(0.263)	(0.362)	(0.313)	(0.186)
Share BA+ (%)	−0.050***	−0.005	−0.015	−0.024	−0.089**	−0.074**	−0.017
	(0.019)	(0.030)	(0.022)	(0.024)	(0.037)	(0.032)	(0.017)
(log) Median income	7.247***	4.887***	6.820***	2.366**	−0.465	−0.495	5.082***
	(1.196)	(0.722)	(0.780)	(0.960)	(1.704)	(1.362)	(0.765)
E. Lubotsky–Wittenberg							
$\sum_j \beta_j$	0.496***	0.274	0.253	0.103	−0.464**	−0.118	−0.061
	(0.044)	(0.192)	(0.166)	(0.220)	(0.203)	(0.348)	(0.187)
Share BA+ (%)	−0.053*	−0.055***	−0.066***	−0.014	−0.079**	−0.013	−0.060***
	(0.028)	(0.015)	(0.020)	(0.026)	(0.033)	(0.029)	(0.019)
(log) Median income	7.169***	1.834***	3.626***	0.666	−0.941	−0.253	3.208***
	(1.868)	(0.688)	(0.945)	(1.292)	(1.723)	(1.274)	(0.841)
F. Yang–Jia–Li (GMM)							
η	2.074***	1.856	1.926**	1.186	−1.295**		
	(0.178)	(1.561)	(0.792)	(2.480)	(0.634)		
Share BA+ (%)	−0.053*	−0.055***	−0.066***	−0.014	−0.079**		
	(0.028)	(0.014)	(0.019)	(0.026)	(0.035)		
(log) Median income	7.169***	1.834***	3.626***	0.666	−0.941		
	(1.831)	(0.673)	(0.919)	(1.268)	(2.108)		
Observations	429	429	429	429	429	429	429

Notes.—Dependent variable is change in Republican vote share in percentage points. Standard errors are clustered by state. $\eta := \widehat{\beta} \times \sigma_H$ is the effect of a one standard deviation increase in the index. There is no recoverable index in the Lubotsky and Wittenberg (2006) approach; instead the sum of the coefficients is reported. J is the number of indicators used to create the indices (see text for further details). Yang–Jia–Li (GMM) estimates are missing in the final two columns due to a failure of the models to converge. $^{***} < 0.01$, $^{**} < 0.05$, $^{*} < 0.1$.

Figure 4 suggests that IV approaches may help resolve the vast discrepancies arising in Table 5, particularly for the other covariates in the model. We consider two IV strategies. First, we use 11 of the 12 formative indicators to create various indices and use the hold out indicator as the instrument. If the hold out variable is a formative indicator (i.e., its weight, λ^* in Equation (6) is non-zero), then the instrument is invalid. However, if $\mathcal{J}$ is large, then the correlation between the hold out indicator and the error term may be quite small. We use dental care as the instrument given its strong correlation with the other observed formative indicators. Second, we follow Blundell et al. (2023) and use the 12 formative indicators to instrument for various indices created from the reflective indicators. The instruments are valid if the formative indicators are orthogonal to the measurement errors in the reflective indicators.

Table 6 presents the results. Columns (2)–(5) display the results where dental care is used to instrument for formative indices excluding dental care. The instrument is fairly strong in all specifications. In addition, the null of exogeneity is rejected in three of four specifications. For comparison (and brevity), the non-IV results can be taken from Column (4) in Table 5, noting that it is not an exact comparison as the indices are created using all 12 formative indicators. Nonetheless, the results suggest a larger positive and statistically significant association between (poor) community health and the change in the two-party Republican vote share in Columns (2)–(4) where exogeneity is rejected. Here, a one standard deviation increase in the index is associated with a 3.7pp to 4.3pp increase in the vote share. The IV results also suggest a larger positive and statistically significant association between median income and the change in the two-party Republican vote share in the same specifications. The association between education and the change in the vote share is never statistically different from zero.

Table 6
Change in Republican Vote Share and Health Indices: IV Results

	Formative Indices				Reflective Indices			
	Mean z-score	PCA	EFA	PLS	Mean z-score	PCA	EFA	PLS
η	3.679^{***}	4.295^{***}	3.901^{***}	2.229^{***}	-1.442^{***}	-1.424^{***}	-1.394^{**}	-1.974^{***}
	(0.775)	(1.016)	(0.801)	(0.304)	(0.539)	(0.537)	(0.546)	(0.350)
Share BA+ (%)	0.038	0.027	0.014	-0.024	-0.105^{***}	-0.105^{***}	-0.103^{***}	-0.074^{***}
	(0.038)	(0.048)	(0.040)	(0.022)	(0.039)	(0.039)	(0.040)	(0.027)
(log) Median	10.633^{***}	13.885^{***}	12.881^{***}	6.677^{***}	-1.537	-1.512	-1.524	-2.258^{**}
income	(1.954)	(3.649)	(2.751)	(0.675)	(1.917)	(1.912)	(1.929)	(1.114)
First-stage F	39.657	47.059	70.152	34.678	50.218	52.658	59.494	13.329
Endog Test	$p = 0.003$	$p = 0.000$	$p = 0.001$	$p = 0.878$	$p = 0.178$	$p = 0.199$	$p = 0.237$	$p = 0.000$
Observations	429	429	429	429	429	429	429	429

Notes.—Dependent variable is change in Republican vote share in percentage points. Standard errors are clustered by state. $\eta := \widehat{\beta} \times \sigma_H$ is the effect of a one standard deviation increase in the index. Columns (2)–(5) report 2SLS estimates where standardized formative indicators (excluding dental care) are instrumented using dental care. Columns (5)–(9) report 2SLS estimates where the standardized reflective health indices are instrumented for using the 12 standardized formative indicators. Endog Test = Wooldridge's robust regression-based test of endogeneity. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Columns (6)–(9) display the results where the 12 formative indicators are used to instrument for reflective indices based on all eight indicators. The instrument is fairly strong in all specifications but the final column. However, the null of exogeneity is not rejected in the three specifications where the instruments are relatively strong. For comparison, the non-IV results are provided in Column (7) in Table 5. Here, the association between (poor) community health and the change in the two-party

Republican vote share continues to be negative and statistically significant, with the point estimates marginally larger in absolute value than the corresponding OLS estimates. The association between education and two-party Republican vote share continues to be negative and statistically significant, while the association between median income and two-party Republican vote share ranges is negative but only statistically significant in the final column where the instruments are relatively weak.

Overall, the results illustrate the sensitivity anticipated by the theory and our simulations, both for the marginal effect of the latent construct itself and other covariates in the model. Index construction is a crucial part of research design and cannot be ignored even when the latent construct itself is not of primary interest. When the analyst changes the measurement model—and thus the source of identifying variation—the implied parameter can change materially, including in sign. The point is not that one column is “right”; rather, absent explicit measurement assumptions, index-based regressions can silently target different objects. In that vein, one possible interpretation of the specific findings here is that a worsening of the observed *inputs* into community health (i.e., the formative indices) is associated with an increase in the two-party Republican vote share, but a worsening of the observed *outputs* of community health (i.e., the reflective indices) is associated with an decrease in the two-party Republican vote share. Moreover, the association between education and median income depends on whether one controls inputs or outputs. This interpretation is consistent with our final recommendation in Section 7.

10 Conclusion

Composite indices are a workhorse of applied economics along with many other disciplines, yet index construction is too often treated as an innocuous data pre-processing step. Our analysis shows this view is untenable. When a latent regressor is replaced by a linear proxy index derived from several manifest variables, the resulting specification embeds identifying assumptions with first-order consequences for estimation and inference. The central lesson is simple but demanding: *index choice is an identification choice*.

Our analysis leads to four core findings. First, replacing a latent regressor with a composite index is not neutral, even when all observed variables are well measured and substantively relevant. Index construction determines how measurement error enters the regression and therefore shapes identification. Second, the consequences of proxy indexing extend beyond the coefficient on the latent construct itself. Proxy indices induce residual confounding (leakage) that can materially distort estimation and inference for other covariates when those covariates are correlated with the index. Third, because the measurement error is nonclassical, estimates can be biased in any direction, generating both Type I and Type II errors, and even sign reversal. Finally, there is no universally valid linear index for regression work. We establish a (near) impossibility result: no linear index can guarantee consistent OLS estimates of all parameters in an otherwise correctly specified multiple regression model. Any index necessarily trades off approximation error in the latent construct against induced correlation with other regressors, making optimality inherently context-dependent.

The empirical exercises underscore these points. Revisiting Ortoleva and Snowberg (2015), we show that substantive conclusions about political behavior are predominantly stable across index approaches, but can still vary meaningfully when the components of the index are entered directly into the regression.

In our original application to the 2024 U.S. presidential election—linking the Congressional District Health Dashboard to precinct-aggregated returns and examining changes in Republican two-party vote share from 2020 to 2024—conclusions are fundamentally altered by measurement choices. These results reinforce the broader message: articulation and justification of a measurement model is not a technical footnote but a core component of research design, deserving the same scrutiny that researchers routinely apply to the choice of instruments, functional forms, and identifying assumptions. Just as Griliches (1985) said.

References

Agostinelli, Francesco and Matthew Wiswall (2025). "Estimating the technology of children's skill formation". *Journal of Political Economy* 133.3, pp. 846–887.

Andersson, Jonas and Jarle Möen (2016). "A simple improvement of the IV-estimator for the classical errors-in-variables problem". *Oxford Bulletin of Economics and Statistics* 78, pp. 113–125.

Angelini, Viola, Marco Bertoni, and Luca Corazzini (2017). "Unpacking the determinants of life satisfaction: a survey experiment". *Journal of the Royal Statistical Society Series A: Statistics in Society* 180.1, pp. 225–246.

Ashenfelter, Orley and Alan Krueger (1994). "Estimates of the economic return to schooling from a new sample of twins". *The American Economic Review* 84.5, pp. 1157–1173.

Ashley, Richard A and Christopher F Parmeter (2015). "When is it justifiable to ignore explanatory variable endogeneity in a regression model?" *Economics Letters* 137, pp. 70–74.

Bickel, Gary et al. (2000). *Guide to Measuring Household Food Security, Revised 2000*. Tech. rep. USDA Food and Nutrition Service Report. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Black, Dan A, Mark C Berger, and Frank A Scott (2000). "Bounding parameter estimates with nonclassical measurement error". *Journal of the American Statistical Association* 95.451, pp. 739–748.

Black, Dan A and Jeffrey A Smith (2006). "Estimating the returns to college quality with multiple proxies for quality". *Journal of Labor Economics* 24.3, pp. 701–728.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez (2013). "Generating skilled self-employment in developing countries: Experimental evidence from Uganda". *The Quarterly Journal of Economics* 129.2, pp. 697–752.

Blau, David M and Donna B Gilleskie (2001). "The effect of health on employment transitions of older men". In: *Worker Well-Being in a Changing Labor Market*. Ed. by Solomon W. Polachek. Bingley, UK: Emerald Group Publishing, pp. 35–65.

Blundell, Richard et al. (2023). "The impact of health on labor supply near retirement". *Journal of Human Resources* 58.1, pp. 282–334.

Bobko, Philip, Philip L Roth, and Maury A Buster (2007). “The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis”. *Organizational Research Methods* 10, pp. 689–709.

Bollen, Kenneth A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons. ISBN: 978-0-471-01171-2.

Bollinger, Christopher R. (2003). “Measurement error in human capital and the Black–White wage gap”. *Review of Economics and Statistics* 85.3, pp. 578–585.

Bollinger, Christopher R. and Jenny Minier (2015). “On the robustness of coefficient estimates to the inclusion of proxy variables”. *Journal of Econometrics* 187.2, pp. 515–525.

Chalak, Karim and Daniel Kim (2024). “Higher order moments for differential measurement error, with application to Tobin’s q and corporate investment”. Available at `https://www.kchalak.com/research`. Accessed 07 April 2025.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi (2012). “Plausibly exogenous”. *The Review of Economics and Statistics* 94.1, pp. 260–272.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010). “Estimating the technology of cognitive and noncognitive skill formation”. *Econometrica* 78.3, pp. 883–931.

Dijkstra, Theo K (2010). “Latent variables and indices: Herman Wold’s basic design and Partial Least Squares”. In: *Handbook of partial least squares*. Ed. by Vincenzo Esposito Vinzi et al. Springer Handbooks of Computational Statistics. Springer.

Dobriban, Edgar (2020). “Permutation methods for factor analysis and PCA”. *The Annals of Statistics* 48.5, pp. 2824–2847.

Dong, Hao and Daniel L. Millimet (2024). “Embrace the noise: It is ok to ignore measurement error in a covariate, sometimes”. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae069.

Edwards, Jeffrey R. and Richard P. Bagozzi (2000). “On the nature and direction of relationships between constructs and measures”. *Psychological Methods* 5.2, pp. 155–174.

Erickson, Timothy and Toni M Whited (2006). “On the accuracy of different measures of q”. *Financial Management* 35.3, pp. 5–33.

Filmer, Deon and Kinnon Scott (2012). “Assessing asset indices”. *Demography* 49, pp. 359–392.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams (2021). “Place-based drivers of mortality: Evidence from migration”. *American Economic Review* 111.8, pp. 2697–2735.

Garmaise, Mark J. (2009). “Ties that truly bind: Noncompetition agreements, executive compensation, and firm investment”. *The Journal of Law, Economics, and Organization* 27.2, pp. 376–425.

Geladi, Paul and Bruce R Kowalski (1986). “Partial least-squares regression: A tutorial”. *Analytica Chimica Acta* 185, pp. 1–17.

Gillen, Ben, Erik Snowberg, and Leeat Yariv (2019). “Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study”. *Journal of Political Economy* 127.4, pp. 1826–1863.

Greco, Salvatore et al. (2018). “Stochastic multi-attribute acceptability analysis (SMAA): An application to the ranking of Italian regions”. *Regional Studies* 52.4, pp. 585–600.

Greco, Salvatore et al. (2019). “On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness”. *Social Indicators Research* 141.1, pp. 61–94.

Griliches, Zvi (1985). “Data and econometricians–The uneasy alliance”. *The American Economic Review* 75.2, pp. 196–200.

— (1986). “Economic data issues”. In: *Handbook of econometrics*. Ed. by Zvi Griliches and Michael Intriligator. Vol. III. North-Holland. Chap. 25, pp. 1465–1514.

Hanushek, Eric A and John E Jackson (1977). “Estimating models with erroneous and unobserved variables”. In: *Statistical methods for social scientists*. Ed. by Eric A Hanushek and John E Jackson. Academic Press, pp. 282–324.

Hotelling, Harold (1933). “Analysis of a complex of statistical variables into principal components”. *Journal of Educational Psychology* 24.6, pp. 417–441.

Hünermund, Paul, Beyers Louw, and Mikko Rönkkö (2025). “The choice of control variables in empirical management research: How causal diagrams can inform the decision”. *The Leadership Quarterly* 36.2, p. 101845.

Jarvis, Cheryl Burke, Scott B Mackenzie, and Philip M Podsakoff (2003). “A critical review of construct indicators and measurement model misspecification in marketing and consumer research”. *Journal of Consumer Research* 30.2, pp. 199–218.

Jolliffe, Ian T. and Jorge Cadima (2016). “Principal component analysis: A review and recent developments”. *Philosophical Transactions of the Royal Society A* 374.2065, p. 20150202.

Jöreskog, Karl G and Arthur S Goldberger (1975). “Estimation of a model with multiple indicators and multiple causes of a single latent variable”. *Journal of the American Statistical Association* 70.351, pp. 631–639.

Jöreskog, Karl G. (1970). “A general method for analysis of covariance structures”. *Biometrika* 57.2, pp. 239–251.

Karagiannis, Giannis and Suzanna-Maria Paleologou (2025). “Governance and economic growth in Africa: Evidence from linear, nonlinear and dynamic panel analysis”. *Empirical Economics* 69.1, pp. 39–75.

Kim, Dongwoo and Daniel Wilhelm (2024). “Powerful t-tests in the presence of nonclassical measurement error”. *Econometric Reviews* 43.6, pp. 345–378.

Kippersluis, Hans van and Cornelius A Rietveld (2018). “Beyond plausibly exogenous”. *The Econometrics Journal* 21.3, pp. 316–331.

Kiviet, Jan F (2016). “When is it really justifiable to ignore explanatory variable endogeneity in a regression model?” *Economics Letters* 145, pp. 192–195.

— (2020). “Testing the impossible: Identifying exclusion restrictions”. *Journal of Econometrics* 218.2, pp. 294–316.

— (2023). “Instrument-free inference under confined regressor endogeneity and mild regularity”. *Econometrics and Statistics* 25, pp. 1–22.

Klein, Roger and Francis Vella (2010). “Estimating a class of triangular simultaneous equations models without exclusion restrictions”. *Journal of Econometrics* 154.2, pp. 154–164.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007). “Experimental analysis of neighborhood effects”. *Econometrica* 75.1, pp. 83–119.

Knaus, Michael C, Michael Lechner, and Anthony Strittmatter (2020). “Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence”. *The Econometrics Journal* 24.1, pp. 134–161.

Kripfganz, Sebastian and Jan F Kiviet (2021). “kinkyreg: Instrument-free inference for linear regression models with endogenous regressors”. *The Stata Journal* 21.3, pp. 772–813.

Krishnakumar, Jaya and A. L. Nagar (2008). “On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models”. *Social Indicators Research* 86.3, pp. 481–496.

Kuncel, Nathan R., Marcus Credé, and Lisa L. Thomas (2005). “The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature”. *Review of Educational Research* 75.1, pp. 63–82.

Levai, Adam and Riccardo Turati (2025). “International immigration and labor regulation”. *The Scandinavian Journal of Economics*. Forthcoming.

Lewbel, Arthur (2012). “Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models”. *Journal of Business & Economic Statistics* 30.1, pp. 67–80.

Lewbel, Arthur, Susanne M Schennach, and Linqi Zhang (2024). “Identification of a triangular two equation system without instruments”. *Journal of Business & Economic Statistics* 42, pp. 14–25.

Lubotsky, Darren and Martin Wittenberg (2006). “Interpretation of regressions with multiple proxies”. *Review of Economics and Statistics* 88.3, pp. 549–562.

Maccini, Sharon and Dean Yang (2009). “Under the weather: Health, schooling, and economic consequences of early-life rainfall”. *American Economic Review* 99.3, pp. 1006–1026.

Mazziotta, Matteo and Adriano Pareto (2016). “On a generalized non-compensatory composite index for measuring socio-economic phenomena”. *Social Indicators Research* 127.3, pp. 983–1003.

— (2018). “Measuring well-being over time: The adjusted Mazziotta–Pareto index versus other non-compensatory indices”. *Social Indicators Research* 136.3, pp. 967–976.

— (2019). “Use and misuse of PCA for measuring well-being”. *Social Indicators Research* 142.2, pp. 451–476.

McCarty, Nolan et al. (2019). “Geography, uncertainty, and polarization”. *Political Science Research and Methods* 7.4, pp. 775–794.

Meijer, Erik, Agnieszka Postepska, and Tom Wansbeek (2025). “Handling multiple proxies”. *Empirical Economics* 69.6, pp. 3063–3087.

Millimet, Daniel L, Ian K McDonough, and Thomas B Fomby (2018). “Financial capability and food security in extremely vulnerable households”. *American Journal of Agricultural Economics* 100, pp. 1224–1249.

Millimet, Daniel L and Christopher F Parmeter (2025). “The impact of measurement error on trends in earnings inequality in the USA”. *Empirical Economics* 69.5, pp. 2727–2753.

Millimet, Daniel L and Travis Whitacre (2025). “Partisan mortality cycles”. *Journal of Population Economics* 38.4, pp. 1–49.

Montero, Eduardo and Dean Yang (2022). “Religious festivals and economic development: Evidence from the timing of Mexican Saint Day festivals”. *American Economic Review* 112.10, pp. 3176–3214.

Mundlak, Yair (1961). “Empirical production function free of management bias”. *Journal of Farm Economics* 43.1, pp. 44–56.

Nevo, Aviv and Adam M Rosen (2012). “Identification with imperfect instruments”. *The Review of Economics and Statistics* 94.3, pp. 659–671.

Ortoleva, Pietro and Erik Snowberg (2015). “Overconfidence in political behavior”. *The American Economic Review* 105.2, pp. 504–535.

Pearson, K. (1901). “On lines and planes of closest fit to systems of points in space”. *Philosophical Magazine* 2.11, pp. 559–572.

Radatz, Laura and Jörg Baten (2025). “Measuring multidimensional inequality and its impact on civil war outbreak in 193 countries, 1810–2010”. *Review of Income and Wealth* 71.2, e70016.

Ravallion, Martin (2012). “Mashup indices of development”. *The World Bank Research Observer* 27.1, pp. 1–32.

Rockwood, Kenneth et al. (2017). “A frailty index based on deficit accumulation quantifies mortality risk in humans and in mice”. *Scientific Reports* 7.1, p. 43068.

Rönkkö, Mikko, Cameron N McIntosh, and John Antonakis (2015). “On the adoption of partial least squares in psychological research: Caveat emptor”. *Personality and Individual Differences* 87, pp. 76–84.

Samuelson, Paul A (1983). *Foundations of economic analysis*. Harvard University Press.

Schennach, Susanne M. (2020). “Chapter 6 - Mismeasured and unobserved variables”. In: *Handbook of Econometrics, Volume 7A*. Ed. by Steven N. Durlauf et al. Vol. 7. Handbook of Econometrics. Elsevier, pp. 487–565.

Solon, Gary (1992). “Intergenerational income mobility in the United States”. *The American Economic Review* 82.3, pp. 393–408.

Spearman, C. (1904). “‘General Intelligence,’ Objectively Determined and Measured”. *The American Journal of Psychology* 15.2, pp. 201–292.

Starr, Evan (2019). “Consider this: Training, wages, and the enforceability of covenants not to compete”. *ILR Review* 72.4, pp. 783–817.

Stoetzer, Lukas F, Xiang Zhou, and Marco Steenbergen (2025). “Causal inference with latent outcomes”. *American Journal of Political Science* 69.2, pp. 624–640.

Thurstone, L. L. (1931). “Multiple-factor analysis”. *Psychological Review* 38.5, pp. 406–427.

Wasfy, Jason H et al. (2020). “Relationship of public health with continued shifting of party voting in the United States”. *Social Science & Medicine* 252, p. 112921.

Woessmann, Ludger (2025). “Skills and earnings: A multidimensional perspective on human capital”. *Annual Review of Economics*.

Wold, Herman (1982). “Soft modeling: The basic design and some extensions”. In: *Systems Under Indirect Observation: Causality, Structure, Prediction*. Ed. by Karl G. Jöreskog and Herman Wold. Vol. 139. Contributions to Economic Analysis. Amsterdam: North-Holland, pp. 1–54.

Wright, Sewall (1921). “Correlation and causation”. *Journal of Agricultural Research* 20.7, pp. 557–585.

Yang, Yimin, Fei Jia, and Haoran Li (2023). “Estimation of panel data models with mixed sampling frequencies”. *Oxford Bulletin of Economics and Statistics* 85.3, pp. 514–544.

Young, Alwyn (2022). “Consistency without inference: Instrumental Variables in practical application”. *European Economic Review* 147, p. 104112.

Zhang, Jeffrey and Junu Lee (2025). “A general condition for bias attenuation by a nondifferentially mismeasured confounder”. *Biometrika*, asaf026.

Rethinking Composite Indices: Reliability, Practical Alternatives, and an Application in Political Economy

Supplemental Appendix

Daniel L. Millimet and Alfredo R. Paloyo

February 26, 2026

Declaration: AI language models (OpenAI's ChatGPT, Google Gemini, and Anthropic's Claude) were used to assist with verification of mathematical derivations and editorial refinement throughout the manuscript but especially in the supplemental appendix. The authors are responsible for the accuracy, validity, and interpretation of all content presented.

A Properties of Linear Proxy Indices in the Reflective-Indicators Model

A.1 Setup

Data-generating process is given by

$$\begin{aligned} y &= \alpha + \beta x^* + \gamma w + \varepsilon \\ z_j &= x^* + u_j, \quad j = 1, ..., J \end{aligned}$$

where $\mathtt{E}[u_j] = 0$, $\mathtt{Var}(u_j) = \sigma^2_{u_j}$, $\mathtt{Cov}(x^*, u_j) = 0$ for all j, $\mathtt{Cov}(u_j, u_k) = 0$ for all $j \neq k$, and $\mathtt{Cov}(x^*, \varepsilon) = 0$.

A.2 Linear Proxy Indices

Standardized Manifest Variables. A generic linear proxy index in terms of the standardized manifest variables is given by

$$x := \sum\nolimits_j \lambda_j \left(\frac{z_j - \overline{z}_j}{\sigma_{z_j}} \right)$$

and the resulting proxy error is

$$\begin{aligned} \mu &:= x - x^* \\ &= \sum\nolimits_j \lambda_j \left(\frac{z_j - \overline{z}_j}{\sigma_{z_j}} \right) - x^* \\ &= \sum\nolimits_j \lambda_j \left(\frac{x^* + u_j - \overline{x}^* - \overline{u}_j}{\sigma_{z_j}} \right) - x^* \\ &= \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) x^* - \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} \right) \overline{x}^* + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \overline{u}_j). \end{aligned}$$

Non-standardized Manifest Variables. A generic linear proxy index in terms of the non-standardized manifest variables is given by

$$x := \sum\nolimits_j \lambda_j z_j$$

and the resulting proxy error is

$$\begin{aligned} \mu &:= x - x^* \\ &= \sum\nolimits_j \lambda_j z_j - x^* \\ &= \sum\nolimits_j \lambda_j (x^* + u_j) - x^* \\ &= \left(\sum\nolimits_j \lambda_j - 1 \right) x^* + \sum\nolimits_j \lambda_j u_j. \end{aligned}$$

A.3 Proxy Error Properties

Setup and Definitions. Structure and notation:

$$\begin{aligned}
z_j &= x^* + u_j, \quad j = 1, \ldots, J \\
x &= \sum\nolimits_j \lambda_j \frac{z_j - \bar{z}_j}{\sigma_{z_j}} \\
q &= \sum\nolimits_j \lambda_j z_j \\
\mu &= x - x^* \\
\eta &= q - x^*
\end{aligned}$$

where $\bar{z}_j$ and $\sigma^2_{z_j}$ are the sample mean and variance of z_j (as defined in Section 2), x (q) is a linear combination of the standardized (non-standardized) z's, and μ (η) is the corresponding proxy error. The expressions derived below treat $\bar{z}_j$ and σ_{z_j} as evaluated at their probability limits; the resulting formulas are the asymptotic approximations reported in Propositions 1 and 2.

Derivation of μ. Note that:

$$\begin{aligned}
\bar{z}_j &= \bar{x}^* + \bar{u}_j \\
z_j - \bar{z}_j &= (x^* - \bar{x}^*) + (u_j - \bar{u}_j)
\end{aligned}$$

Substituting into x:

$$\begin{aligned}
x &= \sum\nolimits_j \lambda_j \frac{(x^* - \bar{x}^*) + (u_j - \bar{u}_j)}{\sigma_{z_j}} \\
&= (x^* - \bar{x}^*) \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j).
\end{aligned}$$

The proxy error is given by:

$$\begin{aligned}
\mu &= (x^* - \bar{x}^*) \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j) - x^* \\
&= (x^* - \bar{x}^*) \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - x^* + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j) \\
&= x^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - \bar{x}^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - x^* + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j) \\
&= x^* \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) - \bar{x}^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} + \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} (u_j - \bar{u}_j).
\end{aligned}$$

Derivation of η_i. For the non-standardized case:

$$\begin{aligned} q &= \sum\nolimits_j \lambda_j z_j = \sum\nolimits_j \lambda_j (x^* + u_j) \\ &= x^* \sum\nolimits_j \lambda_j + \sum\nolimits_j \lambda_j u_j \end{aligned}$$

The proxy error is given by:

$$\eta = x^* \left(\sum\nolimits_j \lambda_j - 1 \right) + \sum\nolimits_j \lambda_j u_j$$

Statistical Properties. Assume $\mathtt{E}[u_j] = 0$, $\mathtt{Var}(u_j) = \sigma_{u_j}^2$, and $\mathtt{Cov}(u_j, u_k) = 0 \; \forall j \neq k$.

- For μ
  - Expected Value:

    Taking expectations:

    $$\mathtt{E}[\mu] = \mathtt{E}[x^*] \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) - \mathtt{E}[\bar{x}^*] \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} + 0$$

    Since $\mathtt{E}[\bar{x}^*] = \mathtt{E}[x^*] = \mathtt{E}[x^*]$:

    $$\begin{aligned} \mathtt{E}[\mu] &= \mathtt{E}[x^*] \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) - \mathtt{E}[x^*] \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} \\ &= \mathtt{E}[x^*] \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - \mathtt{E}[x^*] - \mathtt{E}[x^*] \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} \\ &= -\mathtt{E}[x^*] \end{aligned}$$

  - Variance: For large N, treating $\bar{u}_j \approx 0$ and σ_{z_j} as approximately constant:

    $$\mathtt{Var}(\mu) \approx \mathtt{Var} \left(x^* \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) - \bar{x}^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} \right) + \mathtt{Var} \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} u_j \right)$$

    Since $\mathtt{Var}(\bar{x}^*) = \mathtt{Var}(x^*)/N \approx 0$ for large N:

    $$\mathtt{Var}(\mu) \approx \mathtt{Var}(x^*) \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right)^2 + \sum\nolimits_j \frac{\lambda_j^2}{\sigma_{z_j}^2} \sigma_{u_j}^2$$

  - Covariance with x^*:

    $$\mathtt{Cov}(\mu, x^*) = \mathtt{Cov} \left(x^* \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right) - \bar{x}^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}}, x^* \right)$$

    For large N, $\mathtt{Cov}(\bar{x}^*, x^*) = \mathtt{Var}(x^*)/N \approx 0$:

    $$\mathtt{Cov}(\mu, x^*) = \mathtt{Var}(x^*) \left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1 \right)$$

- Covariance with w:

$$\mathtt{Cov}(\mu, w) = \mathtt{Cov}\left(x^*\left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right) - \bar{x}^* \sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}}, w\right)$$

  For large N, $\mathtt{Cov}(\bar{x}^*, w) = \mathtt{Cov}(x^*, w)/N \approx 0$:

$$\mathtt{Cov}(\mu, w) = \mathtt{Cov}(x^*, w)\left(\sum\nolimits_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right)$$

- For η

  - Expected Value:

$$\mathtt{E}[\eta] = \mathtt{E}[x^*]\left(\sum\nolimits_j \lambda_j - 1\right)$$

  - Variance:

$$\mathtt{Var}(\eta) = \mathtt{Var}(x^*)\left(\sum\nolimits_j \lambda_j - 1\right)^2 + \sum\nolimits_j \lambda_j^2 \sigma_{u_j}^2$$

  - Covariance with x^*:

$$\mathtt{Cov}(\eta, x^*) = \mathtt{Var}(x^*)\left(\sum\nolimits_j \lambda_j - 1\right)$$

  - Covariance with w:

$$\mathtt{Cov}(\eta, w) = \mathtt{Cov}(x^*, w)\left(\sum\nolimits_j \lambda_j - 1\right)$$

Summary Comparison.

Table A.1
Proxy Error Properties

Property	**Standardized Error (μ)**	**Non-Standardized Error (η)**
$\mathtt{E}[\cdot]$	$-\mathtt{E}[x^*]$	$\mathtt{E}[x^*]\left(\sum_j \lambda_j - 1\right)$
$\mathtt{Var}(\cdot)$	$\mathtt{Var}(x^*)\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right)^2 + \sum_j \frac{\lambda_j^2 \sigma_{u_j}^2}{\sigma_{z_j}^2}$	$\mathtt{Var}(x^*)\left(\sum_j \lambda_j - 1\right)^2 + \sum_j \lambda_j^2 \sigma_{u_j}^2$
$\mathtt{Cov}(\cdot, x^*)$	$\mathtt{Var}(x^*)\left(\sum_j \frac{\lambda_j}{\sigma_{z_j}} - 1\right)$	$\mathtt{Var}(x^*)\left(\sum_j \lambda_j - 1\right)$

Table A.2
Conditions for Classical Proxy Errors

Property	**Standardized Error**	**Non-Standardized Error**
Expression for error	Dependent on x^* and u_j	Dependent on x^* and u_j
$\mathtt{E}[\text{error}]$	Non-zero (unless $\mathtt{E}[x^*] = 0$)	Non-zero (unless $\sum_j \lambda_j = 1$)
$\mathtt{Cov}(\text{error}, x^*)$	Non-zero	Non-zero (unless $\sum_j \lambda_j = 1$)

B Properties of Linear Proxy Indices in the Formative-Indicators Model

B.1 Setup

Data-generating process is given by

$$\begin{aligned} y &= \alpha + \beta x^* + \gamma w + \varepsilon \\ x^* &= \sum_{j=1}^{\mathcal{J}} \lambda_j^* z_j \end{aligned}$$

where $\mathtt{Cov}(z_j, \varepsilon) = 0 \; \forall j = 1, ..., \mathcal{J}$.

B.2 Linear Proxy Indices

Standardized Manifest Variables. A generic linear proxy index in terms of $J \leq \mathcal{J}$ standardized manifest variables is given by

$$x := \sum_{j=1}^{J} \lambda_j \left(\frac{z_j - \overline{z}_j}{\sigma_{z_j}} \right)$$

and the resulting proxy error is

$$\begin{aligned} \mu &:= x - x^* \\ &= \sum_{j=1}^{J} \lambda_j \left(\frac{z_j - \overline{z}_j}{\sigma_{z_j}} \right) - \sum_{j=1}^{\mathcal{J}} \lambda_j^* z_j. \end{aligned}$$

Non-standardized Manifest Variables. A generic linear proxy index in terms of $J \leq \mathcal{J}$ non-standardized manifest variables is given by

$$x := \sum_{j=1}^{J} \lambda_j z_j$$

and the resulting proxy error is

$$\begin{aligned} \mu &:= x - x^* \\ &= \sum_{j=1}^{J} (\lambda_j - \lambda_j^*) z_j - \sum_{j=J+1}^{\mathcal{J}} \lambda_j^* z_j. \end{aligned}$$

B.3 Proxy Error Properties

Setup and Definitions. Structure and notation:

$$\begin{aligned} x^* &= \sum\nolimits_j \lambda_j^* z_j, \quad j = 1, \ldots, \mathcal{J} \\ x &= \sum\nolimits_j \lambda_j \frac{z_j - \bar{z}_j}{\sigma_{z_j}}, \quad j = 1, \ldots, J \\ q &= \sum\nolimits_j \lambda_j z_j, \quad j = 1, \ldots, J \\ \mu &= x - x^* \\ \eta &= q - x^* \end{aligned}$$

Derivation of μ. For the standardized case, the proxy error is:

$$\begin{aligned} \mu &= \sum\nolimits_j \lambda_j \frac{z_j - \bar{z}_j}{\sigma_{z_j}} - \sum\nolimits_j \lambda_j^* z_j \\ &= \sum\nolimits_j \lambda_j \frac{z_j}{\sigma_{z_j}} - \sum\nolimits_j \lambda_j \frac{\bar{z}_j}{\sigma_{z_j}} - \sum\nolimits_j \lambda_j^* z_j \\ &= \sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^* \right) z_j - \sum\nolimits_j \lambda_j \frac{\bar{z}_j}{\sigma_{z_j}} \end{aligned}$$

where the summations are over $j = 1, \ldots, \mathcal{J}$ and $\lambda_j = 0$ for $j = J + 1, \ldots, \mathcal{J}$.

Derivation of η. For the non-standardized case, the proxy error is:

$$\begin{aligned} \eta &= \sum\nolimits_j \lambda_j z_j - \sum\nolimits_j \lambda_j^* z_j \\ &= \sum\nolimits_j (\lambda_j - \lambda_j^*) z_j \end{aligned}$$

where the summations are over $j = 1, \ldots, \mathcal{J}$ and $\lambda_j = 0$ for $j = J + 1, \ldots, \mathcal{J}$.

Statistical Properties.

- For μ
  - Expected Value:

$$\begin{aligned} \mathtt{E}[\mu] &= \mathtt{E}\left[\sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^* \right) z_j - \sum\nolimits_j \lambda_j \frac{\bar{z}_j}{\sigma_{z_j}} \right] \\ &= \sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^* \right) \mathtt{E}[z_j] - \sum\nolimits_j \lambda_j \frac{\mathtt{E}[\bar{z}_j]}{\sigma_{z_j}} \end{aligned}$$

Since $\mathtt{E}[\bar{z}_j] = \mathtt{E}[z_j]$:

$$\begin{aligned}\mathtt{E}[\mu] &= \sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) \mathtt{E}[z_j] - \sum\nolimits_j \lambda_j \frac{\mathtt{E}[z_j]}{\sigma_{z_j}} \\ &= -\sum\nolimits_j \lambda_j^* \mathtt{E}[z_j]\end{aligned}$$

- Variance:

$$\begin{aligned}\mathtt{Var}(\mu) &= \mathtt{Var}\left(\sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) z_j\right) \\ &= \sum\nolimits_j \sum\nolimits_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right)\left(\frac{\lambda_k}{\sigma_{z_k}} - \lambda_k^*\right) \mathtt{Cov}(z_j, z_k)\end{aligned}$$

- Covariance with x^*:

$$\begin{aligned}\mathtt{Cov}(\mu, x^*) &= \mathtt{Cov}\left(\sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) z_j, \sum\nolimits_k \lambda_k^* z_k\right) \\ &= \sum\nolimits_j \sum\nolimits_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) \lambda_k^* \mathtt{Cov}(z_j, z_k)\end{aligned}$$

- Covariance with w:

$$\begin{aligned}\mathtt{Cov}(\mu, w) &= \mathtt{Cov}\left(\sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) z_j, w\right) \\ &= \sum\nolimits_j \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right) \mathtt{Cov}(z_j, w)\end{aligned}$$

- For η

  - Expected Value:

$$\begin{aligned}\mathtt{E}[\eta] &= \mathtt{E}\left[\sum\nolimits_j (\lambda_j - \lambda_j^*) z_j\right] \\ &= \sum\nolimits_j (\lambda_j - \lambda_j^*) \mathtt{E}[z_j]\end{aligned}$$

  - Variance:

$$\begin{aligned}\mathtt{Var}(\eta) &= \mathtt{Var}\left(\sum\nolimits_j (\lambda_j - \lambda_j^*) z_j\right) \\ &= \sum\nolimits_j \sum\nolimits_k (\lambda_j - \lambda_j^*)(\lambda_k - \lambda_k^*) \mathtt{Cov}(z_j, z_k)\end{aligned}$$

  - Covariance with x^*:

$$\begin{aligned}\mathtt{Cov}(\eta, x^*) &= \mathtt{Cov}\left(\sum\nolimits_j (\lambda_j - \lambda_j^*) z_j, \sum\nolimits_k \lambda_k^* z_k\right) \\ &= \sum\nolimits_j \sum\nolimits_k (\lambda_j - \lambda_j^*) \lambda_k^* \mathtt{Cov}(z_j, z_k)\end{aligned}$$

- Covariance with w:

$$\begin{aligned}\mathsf{Cov}(\eta, w) &= \mathsf{Cov}\left(\sum\nolimits_j (\lambda_j - \lambda_j^*) z_j, w\right) \\ &= \sum\nolimits_j (\lambda_j - \lambda_j^*) \mathsf{Cov}(z_j, w)\end{aligned}$$

Summary Comparison.

Table B.3
Proxy Error Properties

Property	**Standardized Error (μ)**	**Non-Standardized Error (η)**
Expectation	$-\sum_j \lambda_j^* \mathtt{E}[z_j]$	$\sum_j (\lambda_j - \lambda_j^*)\mathtt{E}[z_j]$
Variance	$\sum_j \sum_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right)\left(\frac{\lambda_k}{\sigma_{z_k}} - \lambda_k^*\right)\mathtt{Cov}(z_j, z_k)$	$\sum_j \sum_k (\lambda_j - \lambda_j^*)(\lambda_k - \lambda_k^*)\mathtt{Cov}(z_j, z_k)$
Covariance	$\sum_j \sum_k \left(\frac{\lambda_j}{\sigma_{z_j}} - \lambda_j^*\right)\lambda_k^*\mathtt{Cov}(z_j, z_k)$	$\sum_j \sum_k (\lambda_j - \lambda_j^*)\lambda_k^*\mathtt{Cov}(z_j, z_k)$

Table B.4
Conditions for Classical Proxy Errors

Property	**Standardized Error**	**Non-Standardized Error**
Expression for error	Dependent on x^*	Dependent on x^*
$\mathtt{E}$[error]	Non-zero (unless $\mathtt{E}[x^*] = 0$)	Non-zero (unless $\lambda_j - \lambda_j^* = 0\ \forall j$)
$\mathtt{Cov}$(error, x^*)	Non-zero	Non-zero (unless $\lambda_j - \lambda_j^* = 0\ \forall j$)

C OLS Properties

C.1 Probability Limit

The regression model using the proxy index x in place of x^* is given by:

$$y = \alpha + \beta x + \gamma w + (\varepsilon - \beta \mu)$$

The OLS estimator of β converges to:

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} &= \frac{\mathtt{Cov}(x, y|w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Cov}(x, x^*|w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Cov}(x, x - \mu \mid w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Var}(x|w) - \mathtt{Cov}(x, \mu \mid w)}{\mathtt{Var}(x|w)} \\
&= \beta\left[1 - \frac{\mathtt{Cov}(x, \mu) - \frac{\mathtt{Cov}(x,w)\mathtt{Cov}(w,\mu)}{\mathtt{Var}(w)}}{\mathtt{Var}(x)(1 - R^2_{x|w})}\right] = \beta\left[1 - \frac{\mathtt{Var}(x)(1 - \delta^*_x) - \delta(\delta - \delta^*)\mathtt{Var}(w)}{\mathtt{Var}(x)(1 - R^2_{x|w})}\right] \\
&= \beta\left[\frac{\mathtt{Var}(x)(1 - R^2_{x|w}) - \mathtt{Var}(x)(1 - \delta^*_x) + \delta(\delta - \delta^*)\mathtt{Var}(w)}{\mathtt{Var}(x)(1 - R^2_{x|w})}\right] \\
&= \beta\left[\frac{\mathtt{Var}(x) - \delta^2\mathtt{Var}(w) - \mathtt{Var}(x) + \delta^*_x\mathtt{Var}(x) + \delta^2\mathtt{Var}(w) - \delta\delta^*\mathtt{Var}(w)}{\mathtt{Var}(x)(1 - R^2_{x|w})}\right] \\
&= \beta\left[\frac{\delta^*_x\mathtt{Var}(x) - \delta\delta^*\mathtt{Var}(w)}{\mathtt{Var}(x)(1 - R^2_{x|w})}\right] = \beta\left[\frac{\delta^*_x - \delta\delta^*\frac{\mathtt{Var}(w)}{\mathtt{Var}(x)}}{1 - R^2_{x|w}}\right] = \beta\left[\frac{\delta^*_x - \frac{\delta^*}{\delta}R^2_{x|w}}{1 - R^2_{x|w}}\right]
\end{aligned}$$

where $\delta^* := \mathtt{Cov}(x^*, w)/\mathtt{Var}(w)$ is the OLS coefficient from the regression of x^* on w, $\delta := \mathtt{Cov}(x, w)/\mathtt{Var}(w)$ is the OLS coefficient from the regression of x on w, and $\delta^*_x := \mathtt{Cov}(x^*, x)/\mathtt{Var}(x)$ is the OLS coefficient from the regression of x^* on x. If $\mathtt{Cov}(x^*, w) = 0$ (reflective model) or $\mathtt{Cov}(z_j, w) = 0\ \forall j$ (formative model), then $\mathtt{Cov}(\mu, w) = 0$ and this simplifies to

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta\delta^*_x.$$

Alternatively,

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} &= \frac{\mathtt{Cov}(x, y|w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Cov}(x, x^*|w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Cov}(x, x - \mu \mid w)}{\mathtt{Var}(x|w)} = \beta\frac{\mathtt{Var}(x|w) - \mathtt{Cov}(x, \mu \mid w)}{\mathtt{Var}(x|w)} \\
&= \beta\left[\frac{\mathtt{Var}(x|w) - \mathtt{Cov}(x^*, \mu \mid w) - \mathtt{Var}(\mu \mid w)}{\mathtt{Var}(x|w)}\right] = \beta\left[\frac{\mathtt{Var}(x^*|w) - \mathtt{Cov}(x^*, \mu \mid w)}{\mathtt{Var}(x|w)}\right] \\
&= \beta\left[\frac{\mathtt{Var}(x^*)\left(1 - R^2_{x^*|w}\right)}{\mathtt{Var}(x)\left(1 - R^2_{x|w}\right)} + \frac{\mathtt{Cov}(x^*, \mu) - \mathtt{Cov}(x^*, w)\mathtt{Cov}(w, \mu)/\mathtt{Var}(w)}{\mathtt{Var}(x)\left(1 - R^2_{x|w}\right)}\right] \\
&= \beta\left[\frac{\left(1 - R^2_{x^*|w}\right) + \mathtt{Cov}(x^*, \mu) - \delta^*\mathtt{Cov}(w, \mu)}{\mathtt{Var}(x)\left(1 - R^2_{x|w}\right)}\right].
\end{aligned}$$

The OLS estimator of γ converges to:

$$
\begin{aligned}
\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} &= \frac{\mathtt{Cov}(w,y\mid x)}{\mathtt{Var}(w\mid x)} = \frac{\beta\mathtt{Cov}(w,x^{*}\mid x)+\gamma\mathtt{Var}(w\mid x)}{\mathtt{Var}(w\mid x)} \\
&= \frac{\beta\mathtt{Cov}(w,x-\mu\mid x)+\gamma\mathtt{Var}(w\mid x)}{\mathtt{Var}(w\mid x)} = \frac{-\beta\mathtt{Cov}(w,\mu\mid x)+\gamma\mathtt{Var}(w\mid x)}{\mathtt{Var}(w\mid x)} \\
&= \gamma-\beta\left[\frac{\mathtt{Cov}(w,\mu\mid x)}{\mathtt{Var}(w\mid x)}\right] = \gamma-\beta\left[\frac{\mathtt{Cov}(w,\mu)-\mathtt{Cov}(w,x)\mathtt{Cov}(\mu,x)/\mathtt{Var}(x)}{\mathtt{Var}(w)\left(1-R^{2}_{x|w}\right)}\right] \\
&= \gamma-\beta\left[\frac{\mathtt{Cov}(w,\mu)\mathtt{Var}(x)/\mathtt{Var}(w)-\delta\mathtt{Cov}(\mu,x)}{\mathtt{Var}(x)\left(1-R^{2}_{x|w}\right)}\right] \\
&= \gamma+\beta\delta\left[\frac{\mathtt{Cov}(\mu,x)-\mathtt{Cov}(w,\mu)\mathtt{Var}(x)/\delta\mathtt{Var}(w)}{\mathtt{Var}(x)\left(1-R^{2}_{x|w}\right)}\right] \\
&= \gamma+\beta\delta\left[\frac{\mathtt{Var}(x)-\mathtt{Cov}(x,x^{*})-\frac{\mathtt{Cov}(w,x)-\mathtt{Cov}(w,x^{*})}{\mathtt{Cov}(x,w)}\mathtt{Var}(x)}{\mathtt{Var}(x)\left(1-R^{2}_{x|w}\right)}\right] \\
&= \gamma+\beta\delta\left[\frac{\mathtt{Var}(x)-\mathtt{Cov}(x,x^{*})-\left(1-\frac{\mathtt{Cov}(w,x^{*})}{\mathtt{Cov}(x,w)}\right)\mathtt{Var}(x)}{\mathtt{Var}(x)\left(1-R^{2}_{x|w}\right)}\right] \\
&= \gamma+\beta\delta\left[\frac{-\mathtt{Cov}(x,x^{*})+\left(\frac{\mathtt{Cov}(w,x^{*})}{\mathtt{Cov}(x,w)}\right)\mathtt{Var}(x)}{\mathtt{Var}(x)\left(1-R^{2}_{x|w}\right)}\right] = \gamma+\beta\delta\left[\frac{\frac{\mathtt{Cov}(w,x^{*})}{\mathtt{Cov}(x,w)}-\delta^{*}_{x}}{1-R^{2}_{x|w}}\right] \\
&= \gamma+\beta\delta\left[\frac{\frac{\delta^{*}}{\delta}-\delta^{*}_{x}}{1-R^{2}_{x|w}}\right] = \gamma+\beta\delta\left[\frac{\frac{\delta^{*}}{\delta}-\delta^{*}_{x}+\frac{\delta^{*}}{\delta}R^{2}_{x|w}-\frac{\delta^{*}}{\delta}R^{2}_{x|w}}{1-R^{2}_{x|w}}\right] \\
&= \gamma+\beta\delta\left[\frac{\frac{\delta^{*}}{\delta}\left(1-R^{2}_{x|w}\right)-\delta^{*}_{x}+\frac{\delta^{*}}{\delta}R^{2}_{x|w}}{1-R^{2}_{x|w}}\right] = \gamma+\beta\delta\left[\frac{\delta^{*}}{\delta}-\frac{\delta^{*}_{x}-\frac{\delta^{*}}{\delta}R^{2}_{x|w}}{1-R^{2}_{x|w}}\right] \\
&= \gamma+\delta\left[\beta\frac{\delta^{*}}{\delta}-\mathtt{plim}\,\widehat{\beta}\right].
\end{aligned}
$$

Alternatively,

$$\begin{aligned}
\texttt{plim}\,\widehat{\gamma}^{\text{OLS}} &= \frac{\texttt{Cov}(w,y \mid x)}{\texttt{Var}(w \mid x)} = \frac{\beta\texttt{Cov}(w,x^* \mid x) + \gamma\texttt{Var}(w \mid x)}{\texttt{Var}(w \mid x)} = \frac{\beta\texttt{Cov}(w,x-\mu \mid x) + \gamma\texttt{Var}(w \mid x)}{\texttt{Var}(w \mid x)} \\
&= \frac{-\beta\texttt{Cov}(w,\mu \mid x) + \gamma\texttt{Var}(w \mid x)}{\texttt{Var}(w \mid x)} = \gamma - \beta\left[\frac{\texttt{Cov}(w,\mu \mid x)}{\texttt{Var}(w \mid x)}\right] \\
&= \gamma - \beta\left[\frac{\texttt{Cov}(w,\mu) - \texttt{Cov}(w,x)\texttt{Cov}(\mu,x)/\texttt{Var}(x)}{\texttt{Var}(w)\left(1-R^2_{x|w}\right)}\right] \\
&= \gamma - \beta\left[\frac{\texttt{Cov}(w,\mu)\texttt{Var}(x)/\texttt{Var}(w) - \delta\texttt{Cov}(\mu,x)}{\texttt{Var}(x)\left(1-R^2_{x|w}\right)}\right] \\
&= \gamma + \beta\delta\left[\frac{\texttt{Cov}(\mu,x) - \texttt{Cov}(w,\mu)\texttt{Var}(x)/\delta\texttt{Var}(w)}{\texttt{Var}(x)\left(1-R^2_{x|w}\right)}\right] \\
&= \gamma + \beta\delta\left[\frac{\texttt{Cov}(x^*,\mu) + \texttt{Var}(\mu) - \frac{\texttt{Cov}(w,\mu)}{\delta}\frac{\texttt{Var}(x)}{\texttt{Var}(w)}}{\texttt{Var}(x)\left(1-R^2_{x|w}\right)}\right].
\end{aligned}$$

C.2 The Statistical Requirement for Simultaneous Consistency

The OLS estimate of β converges to

$$\texttt{plim}\,\widehat{\beta}^{\text{OLS}} = \beta\left[\frac{\texttt{Var}(x^*|w) + \texttt{Cov}(x^*,\mu \mid w)}{\texttt{Var}(x|w)}\right]$$

Using the identity $\texttt{Var}(x|w) = \texttt{Var}(x^*|w) + \texttt{Var}(\mu \mid w) + 2\texttt{Cov}(x^*,\mu \mid w)$, the condition for consistency of $\widehat{\beta}$ simplifies to

$$\texttt{Cov}(x,\mu \mid w) = 0 \implies \texttt{Cov}(x,\mu) = \frac{\texttt{Cov}(x,w)\texttt{Cov}(w,\mu)}{\texttt{Var}(w)}$$

The OLS estimate of γ converges to

$$\texttt{plim}\,\widehat{\gamma}^{\text{OLS}} = \gamma + \beta\frac{\texttt{Cov}(x^*,w \mid x)}{\texttt{Var}(w \mid x)}$$

Consistency of $\widehat{\gamma}$ requires $\texttt{Cov}(x^*,w \mid x) = 0$. Since $x^* = x - \mu$ and x is constant relative to the conditioning set, this requirement is equivalent to $\texttt{Cov}(w,\mu \mid x) = 0$, where

$$\texttt{Cov}(w,\mu \mid x) = 0 \implies \texttt{Cov}(w,\mu) = \frac{\texttt{Cov}(w,x)\texttt{Cov}(x,\mu)}{\texttt{Var}(x)}$$

If both conditions hold simultaneously, substitution of $\mathtt{Cov}(x, \mu)$ into the second requirement yields:

$$\begin{aligned}\mathtt{Cov}(w, \mu) &= \mathtt{Cov}(w, \mu)\left[\frac{(\mathtt{Cov}(x, w))^2}{\mathtt{Var}(w)\mathtt{Var}(x)}\right] \\ &= \mathtt{Cov}(w, \mu)R^2_{x|w}\end{aligned}$$

Excluding the case of perfect multicollinearity ($R^2_{x|w} = 1$), this equality requires $\mathtt{Cov}(w, \mu) = 0$. From the first condition, $\mathtt{Cov}(w, \mu) = 0$ implies $\mathtt{Cov}(x, \mu) = 0$. Expanding this as $\mathtt{Cov}(x^* + \mu, \mu) = 0$, we arrive at

$$\mathtt{Cov}(x^*, \mu) = -\mathtt{Var}(\mu).$$

This characterizes the mean-reverting measurement error described in condition (b) of the Corollary.

C.3 Optimal Index Weights and the Structural Requirement

The structural model is given by

$$y = \alpha + \beta x^* + \gamma w + \varepsilon$$

where x^* is related to the observed indicators $\mathbf{z}$ and w through the linear projection:

$$x^* = \mathbf{z}\boldsymbol{\delta}^* + \zeta w + \nu,$$

with $\mathtt{Cov}(\mathbf{z}, \nu) = \mathtt{Cov}(w, \nu) = 0$. The linear index $x = \mathbf{z}\boldsymbol{\lambda}$ is used to proxy for x^* and OLS is used to estimate

$$y = \alpha + \beta x + \gamma w + (\varepsilon - \beta\mu),$$

where $\mu = x - x^*$. Assuming $\beta \neq 0$ and $\mathtt{Var}(w \mid \mathbf{z}) > 0$, the probability limits of the OLS estimators are:

$$\begin{aligned}\mathtt{plim}\widehat{\beta} &= \beta\frac{\mathtt{Cov}(x^*, x \mid w)}{\mathtt{Var}(x \mid w)} \\ \mathtt{plim}\widehat{\gamma} &= \gamma + \beta\frac{\mathtt{Cov}(x^*, w \mid x)}{\mathtt{Var}(w \mid x)}\end{aligned}$$

Consistency of $\widehat{\beta}$ requires $\mathtt{Cov}(x^*, x \mid w) = \mathtt{Var}(x \mid w)$. Substituting the definitions of x^* and x, and noting that $\mathtt{Cov}(w, \mathbf{z}\boldsymbol{\lambda} \mid w) = 0$ and $\mathtt{Cov}(\nu, \mathbf{z}\boldsymbol{\lambda} \mid w) = 0$, this condition is satisfied if $\boldsymbol{\lambda} = \boldsymbol{\delta}^*$.

For simultaneous consistency, $\widehat{\gamma}$ also requires $\mathtt{Cov}(x^*, w \mid x) = 0$. Substituting the linear projection for x^*:

$$\mathtt{Cov}(\mathbf{z}\boldsymbol{\delta}^* + \zeta w + \nu, w \mid \mathbf{z}\boldsymbol{\lambda}) = 0$$

Let $w = \mathbf{z}\boldsymbol{\pi} + \chi$ be the linear projection of w on $\mathbf{z}$, where $\mathtt{Cov}(\mathbf{z}, \chi) = 0$. Assuming $\mathtt{Cov}(\nu, w \mid \mathbf{z}) = 0$, and imposing the requirement for $\widehat{\beta}$ consistency ($\boldsymbol{\lambda} = \boldsymbol{\delta}^*$), the requirement for $\widehat{\gamma}$ consistency simplifies to:

$$\zeta\,[\mathtt{Var}(\mathbf{z}\boldsymbol{\pi} \mid \mathbf{z}\boldsymbol{\delta}^{*}) + \mathtt{Var}(\chi)] = 0$$

Since $\mathtt{Var}(\chi) = \mathtt{Var}(w \mid \mathbf{z}) > 0$ as long as w is not perfectly collinear with the indicator vector $\mathbf{z}$, simultaneous consistency holds if and only if $\zeta = 0$. This establishes condition (a) of the Corollary.

D PCA in Detail

D.1 Two Manifest Variables ($J = 2$)

1. Formulate the Covariance Matrix: Given $z_1 = x^* + u_1$ and $z_2 = x^* + u_2$, the covariance matrix Σ is:

$$\Sigma = \begin{pmatrix} \mathtt{Var}(z_1) & \mathtt{Cov}(z_1, z_2) \\ \mathtt{Cov}(z_2, z_1) & \mathtt{Var}(z_2) \end{pmatrix} = \begin{pmatrix} 1 + \mathtt{Var}(u_1) & 1 \\ 1 & 1 + \mathtt{Var}(u_2) \end{pmatrix}$$

2. Set up the Characteristic Equation: The eigenvalues φ are found by solving the characteristic equation $\mathtt{det}(\Sigma - \varphi I) = 0$:

$$\begin{aligned} \begin{vmatrix} 1 + \mathtt{Var}(u_1) - \varphi & 1 \\ 1 & 1 + \mathtt{Var}(u_2) - \varphi \end{vmatrix} &= 0 \\ (1 + \mathtt{Var}(u_1) - \varphi)(1 + \mathtt{Var}(u_2) - \varphi) - 1 &= 0 \\ \varphi^2 - (2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2))\varphi + \mathtt{Var}(u_1) + \mathtt{Var}(u_2) + \mathtt{Var}(u_1)\mathtt{Var}(u_2) &= 0 \end{aligned}$$

3. Find the Eigenvalues: Using the quadratic formula, the eigenvalues are:

$$\begin{aligned} \varphi &= \frac{2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2) \pm \sqrt{[2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2)]^2 - 4[\mathtt{Var}(u_1) + \mathtt{Var}(u_2) + \mathtt{Var}(u_1)\mathtt{Var}(u_2)]}}{2} \\ &= \frac{2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2) \pm \sqrt{[\mathtt{Var}(u_1) - \mathtt{Var}(u_2)]^2 + 4}}{2} \end{aligned}$$

The first principal component corresponds to the eigenvector associated with the larger eigenvalue:

$$\varphi_1 = \frac{(2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2)) + \sqrt{(\mathtt{Var}(u_1) - \mathtt{Var}(u_2))^2 + 4}}{2}$$

4. Find the Eigenvector: For the eigenvalue φ_1, we solve $(\Sigma - \varphi_1 I)v = 0$, where $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ is the eigenvector.

$$\begin{pmatrix} 1 + \mathtt{Var}(u_1) - \varphi_1 & 1 \\ 1 & 1 + \mathtt{Var}(u_2) - \varphi_1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

From the first row: $([1 + \mathtt{Var}(u_1) - \varphi_1]v_1 + v_2 = 0 \implies v_2 = [\varphi_1 - 1 - \mathtt{Var}(u_1)]v_1$ Substituting the expression for φ_1:

$$\begin{aligned} v_2 &= \left[\frac{2 + \mathtt{Var}(u_1) + \mathtt{Var}(u_2) + \sqrt{[\mathtt{Var}(u_1) - \mathtt{Var}(u_2)]^2 + 4}}{2} - 1 - \mathtt{Var}(u_1) \right] v_1 \\ &= \left[\frac{\mathtt{Var}(u_2) - \mathtt{Var}(u_1) + \sqrt{[\mathtt{Var}(u_1) - \mathtt{Var}(u_2)]^2 + 4}}{2} \right] v_1 \end{aligned}$$

The eigenvector is proportional to

$$\begin{pmatrix} 1 \\ \frac{\mathtt{Var}(u_2)-\mathtt{Var}(u_1)+\sqrt{[\mathtt{Var}(u_1)-\mathtt{Var}(u_2)]^2+4}}{2} \end{pmatrix}.$$

5. Normalize the Eigenvector: Letting

$$k := \frac{\mathtt{Var}(u_2) - \mathtt{Var}(u_1) + \sqrt{[\mathtt{Var}(u_1) - \mathtt{Var}(u_2)]^2 + 4}}{2},$$

the eigenvector is $\begin{pmatrix} 1 \\ k \end{pmatrix}$. The normalized eigenvector (weights for the first principal component) is:

$$w = \begin{pmatrix} \frac{1}{\sqrt{1+k^2}} \\ \frac{k}{\sqrt{1+k^2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\left(\frac{\mathtt{Var}(u_2)-\mathtt{Var}(u_1)+\sqrt{(\mathtt{Var}(u_1)-\mathtt{Var}(u_2))^2+4}}{2}\right)^2}} \\ \frac{\frac{\mathtt{Var}(u_2)-\mathtt{Var}(u_1)+\sqrt{(\mathtt{Var}(u_1)-\mathtt{Var}(u_2))^2+4}}{2}}{\sqrt{1+\left(\frac{\mathtt{Var}(u_2)-\mathtt{Var}(u_1)+\sqrt{(\mathtt{Var}(u_1)-\mathtt{Var}(u_2))^2+4}}{2}\right)^2}} \end{pmatrix}.$$

D.2 Explicit Calculation for $\mathtt{Var}(u_1) = 0.5$ and $\mathtt{Var}(u_2) = 4$

Non-standardized Manifest Variables. The covariance matrix is $\Sigma = \begin{pmatrix} 1.5 & 1 \\ 1 & 5 \end{pmatrix}$. The characteristic equation is $\varphi^2 - 6.5\varphi + 6.5 = 0$. The eigenvalues are $\varphi_{1,2} = \frac{6.5\pm\sqrt{16.25}}{2} = \frac{6.5\pm 4.0311}{2} \implies \varphi_1 = 5.26555$, $\varphi_2 = 1.23445$. For $\varphi_1 = 5.26555$: $(-3.76555)v_1 + v_2 = 0 \implies v_2 = 3.76555v_1$. Normalized eigenvector: $v_1^2 + v_2^2 = 1 \implies v_1^2 + (3.76555v_1)^2 = 1 \implies v_1^2(1 + 14.180) = 1 \implies v_1^2 = \frac{1}{15.180} \implies v_1 = \pm 0.2566$ $v_2 = \pm 3.76555 \times 0.2566 = \pm 0.9668$. Thus, the weights for the first principal component are approximately $\begin{pmatrix} 0.2566 \\ 0.9668 \end{pmatrix}$.

Standardized Manifest Variables. The correlation matrix is:

$$\rho = \begin{pmatrix} 1 & \frac{1}{\sqrt{7.5}} \\ \frac{1}{\sqrt{7.5}} & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & 0.3651 \\ 0.3651 & 1 \end{pmatrix}$$

The characteristic equation is $(1 - \varphi)^2 - 0.3651^2 = 0$. $\varphi_1 = 1.3651$, $\varphi_2 = 0.6349$. For $\varphi_1 = 1.3651$: $(1 - 1.3651)v_1 + 0.3651v_2 = 0 \implies -0.3651v_1 + 0.3651v_2 = 0 \implies v_1 = v_2$. Normalized eigenvector: $v_1^2 + v_2^2 = 1 \implies v_1^2 + v_1^2 = 1 \implies 2v_1^2 = 1 \implies v_1 = \pm\frac{1}{\sqrt{2}} \approx \pm 0.7071$. The weights for the first principal component after standardization are approximately $\begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix}$.

D.3 Weights on Non-standardized Manifest Variables after PCA Using Standardized Manifest Variables

When PCA is performed on the standardized variables (denoted by Z_1 and Z_2), the first principal component is given by:

$$x^{\texttt{PCA}} = v_1 Z_1 + v_2 Z_2$$

where v_1 and v_2 are the weights obtained from the eigenvector of the correlation matrix. The weights are approximately $v_1 = 0.7071$ and $v_2 = 0.7071$.

The standardized variables are defined as:

$$Z_1 = \frac{z_1 - \texttt{E}[z_1]}{\sqrt{\texttt{Var}(z_1)}}$$

$$Z_2 = \frac{z_2 - \texttt{E}[z_2]}{\sqrt{\texttt{Var}(z_2)}}$$

Substituting these into the expression for the first principal component:

$$x^{\texttt{PCA}} = v_1 \frac{z_1}{\sqrt{\texttt{Var}(z_1)}} + v_2 \frac{z_2}{\sqrt{\texttt{Var}(z_2)}} + C,$$

where $C := -\left[v_1 \texttt{E}[z_1]/\sqrt{\texttt{Var}(z_1)}\right] - \left[v_2 \texttt{E}[z_2]/\sqrt{\texttt{Var}(z_2)}\right]$ Given $\texttt{Var}(z_1) = 1.5$, $\texttt{Var}(z_2) = 5$, and $v_1 = v_2 \approx 0.7071$, the first principal component in terms of the original variables z_1 and z_2 is:

$$\begin{aligned} x^{\texttt{PCA}} &\approx 0.7071 \frac{z_1}{\sqrt{1.5}} + 0.7071 \frac{z_2}{\sqrt{5}} + C \\ &\approx \frac{0.7071}{1.2247} z_1 + \frac{0.7071}{2.2361} z_2 + C \\ &\approx 0.5774 z_1 + 0.3162 z_2 + C \end{aligned}$$

The corresponding weights on the original z_1 and z_2 are approximately 0.5774 and 0.3162, respectively.

D.4 Properties of PCA Weights

Sum of the Weights Using the Covariance Matrix

Let $v = [v_1, \ldots, v_J]'$ be the first principal component (eigenvector) of a $J \times J$ covariance matrix $\mathbf{\Sigma_{zz}}$. Let S_v be the sum of its elements, $S_v := \sum_j v_j$. With PCA the eigenvector is normalized to unit length, $\sum_j v_j^2 = 1$. Assuming that the pairwise covariances $\sigma_{jk} > 0 \; \forall j, k$, then based on the Perron-Frobenius theorem $v_j > 0 \; \forall j$.

To derive the lower bound on S_v, we square S_v:

$$S_v^2 = \left(\sum_j v_j\right)^2 = \sum_j v_j^2 + \sum_{j \neq k} v_j v_k = \sum_j v_j^2 + 2 \sum_{j<k} v_j v_k$$

Given the normalization condition $\sum v_j^2 = 1$, this becomes

$$S_v^2 = 1 + 2\sum\nolimits_{j<k} v_j v_k,$$

where every cross-product term $v_j v_k$ is strictly positive. As a result, $S_v^2 > 1$ if $J \geq 2$. Finally, since S_v is a sum of positive numbers, S_v is the positive root of S_v^2 and $S_v > 1$.

To derive the upper bound, we use the Cauchy–Schwarz inequality:

$$\left(\sum\nolimits_j a_j b_j\right)^2 \leq \left(\sum\nolimits_j a_j^2\right)\left(\sum\nolimits_j b_j^2\right)$$

Let $a_j = v_j$ and $b_j = 1$. Applying the inequality yields:

$$S_v^2 = \left(\sum\nolimits_j v_j \cdot 1\right)^2 \leq \left(\sum\nolimits_j v_j^2\right)\left(\sum\nolimits_j 1^2\right).$$

We can simplify this expression using the fact that $\left(\sum_{j=1}^J 1^2\right) = J$ and $\sum_j v_j^2 = 1$ with PCA. Substitution gives $S_v^2 \leq (1) \cdot (J) = J \implies S_v \leq \sqrt{J}$.

Sum of the Scaled Weights Using the Correlation Matrix

Let v_j be the PCA weights based on the correlation matrix R and define $S_v := \sum_j (v_j/\sigma_{z_j})$. As before, with PCA the eigenvector is normalized to unit length, $\sum_j v_j^2 = 1$. Assuming that the pairwise correlations $r_{jk} > 0\ \forall j, k$, then based on the Perron-Frobenius theorem $v_j > 0\ \forall j$.

To derive the lower bound on S_v, let $\sigma_{\mathtt{max}} := \max(\sigma_{z_1}, ..., \sigma_{z_J})$ be the largest standard deviation among the non-standardized manifest variables. Given the definition of $\sigma_{\mathtt{max}}$, $\sigma_{z_j} \leq \sigma_{\mathtt{max}}\ \forall j \implies \frac{1}{\sigma_{z_j}} \geq \frac{1}{\sigma_{\mathtt{max}}}\ \forall j$. Using this inequality,

$$S_v = \sum\nolimits_j \frac{v_j}{\sigma_{z_j}} \geq \left(\frac{1}{\sigma_{\mathtt{max}}}\right)\sum\nolimits_j v_j$$

Using the same proof as above, $\sum_j v_j > 1$ which implies

$$S_v \geq \left(\frac{1}{\sigma_{\mathtt{max}}}\right)\sum\nolimits_j v_j > \frac{1}{\sigma_{\mathtt{max}}}$$

Thus, $S_v > 1/\sigma_{\mathtt{max}}$.

To derive the upper bound on S_v, we again use the Cauchy–Schwarz inequality:

$$\left(\sum\nolimits_j a_j b_j\right)^2 \leq \left(\sum\nolimits_j a_j^2\right)\left(\sum\nolimits_j b_j^2\right)$$

Let $a_j = v_j$ and $b_j = 1/\sigma_{z_j}$. Applying the inequality yields:

$$S_v^2 = \left(\sum\nolimits_j v_j \cdot \frac{1}{\sigma_{z_j}}\right)^2 \leq \left(\sum\nolimits_j v_j^2\right)\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}^2}\right).$$

Given that $\sum_j v_j^2 = 1$ with PCA, we have $S_v^2 \leq (1) \cdot (\sum_j 1/\sigma_{z_j}^2)$, or

$$S_v \leq \sqrt{\sum_j \frac{1}{\sigma_{z_j}^2}}.$$

D.5 OLS Bias Using PCA Indices

Reflective-Indicators Model

Standardized Manifest Variables. Let $z_j = x^* + u_j$, $j = 1, ..., J$, where x^* is a latent variable and u_j are classical measurement errors with $\mathtt{E}[u_j] = 0$, $\mathtt{Var}(u_j) = \sigma_j^2$, $\mathtt{Cov}(u_j, u_k) = 0\ \forall j \neq k$, and $\mathtt{Cov}(x^*, u_j) = 0\ \forall j$. Let v be the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the standardized z's. The standardized z's are given by $Z_j = \frac{z_j - \mathtt{E}[z_j]}{\sigma_{z_j}}$. The index x is the first principal component, $x := \sum_j v_j Z_j$. The regression model is

$$y = \alpha + \beta x^* + \gamma w + \varepsilon.$$

From Section C.1 the plim of the OLS estimates replacing x^* with x is given by:

$$\begin{aligned} \mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} &= \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2} \right] \\ \mathtt{plim}\,\widehat{\gamma}^{\mathrm{OLS}} &= \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2} \right]. \end{aligned}$$

Since $x := \sum_j v_j Z_j$, where $Z_j := (x^* + u_j - \mathtt{E}[x^*])/\sigma_{z_j}$ denotes the standardized manifest variable j, and $\mathtt{E}[u_j] = 0$, we can express x as

$$\begin{aligned} x &= \sum_j v_j \left(\frac{x^* + u_j}{\sigma_{z_j}} \right) - \sum_j v_j \left(\frac{\mathtt{E}[x^*]}{\sigma_{z_j}} \right) = \sum_j \left(\frac{v_j}{\sigma_{z_j}} \right) x^* + \sum_j \left(\frac{v_j}{\sigma_{z_j}} \right) u_j - \sum_j v_j \left(\frac{\mathtt{E}[x^*]}{\sigma_{z_j}} \right) \\ &= \sum_j c_j x^* + \sum_j c_j u_j - \sum_j c_j \mathtt{E}[x^*], \end{aligned}$$

where $c_j := v_j/\sigma_{z_j}$. The proxy error is

$$\mu = \left(\sum_j c_j - 1 \right) x^* + \sum_j c_j u_j - \mathtt{E}[x^*] \sum_j c_j.$$

The covariance terms are:

$$\begin{aligned}
\texttt{Cov}(x^*,\mu) &= \texttt{Cov}\left(x^*, \left(\sum\nolimits_j c_j - 1\right) x^* + \sum\nolimits_j c_j u_j\right) = \left(\sum\nolimits_j c_j - 1\right) \texttt{Var}(x^*) + \sum\nolimits_j c_j \texttt{Cov}(x^*, u_j) \\
&= \left(\sum\nolimits_j c_j - 1\right) \texttt{Var}(x^*) \\
\texttt{Var}(x) &= \texttt{Var}\left(\sum\nolimits_j c_j x^* + \sum\nolimits_j c_j u_j\right) = \left(\sum\nolimits_j c_j\right)^2 \texttt{Var}(x^*) + \sum\nolimits_j c_j^2 \texttt{Var}(u_j) \\
\texttt{Cov}(w,\mu) &= \texttt{Cov}\left(w, \left(\sum\nolimits_j c_j - 1\right) x^* + \sum\nolimits_j c_j u_j\right) = \left(\sum\nolimits_j c_j - 1\right) \texttt{Cov}(w, x^*) + \sum\nolimits_j c_j \texttt{Cov}(w, u_j) \\
&= \left(\sum\nolimits_j c_j - 1\right) \texttt{Cov}(w, x^*) \\
\texttt{Cov}(x, w) &= \left(\sum\nolimits_j c_j\right) \texttt{Cov}(w, x^*) \\
\texttt{Cov}(x^*, x) &= \left(\sum\nolimits_j c_j\right) \texttt{Var}(x^*)
\end{aligned}$$

The regression coefficients:

$$\begin{aligned}
\delta_x^* &= \frac{\texttt{Cov}(x^*, x)}{\texttt{Var}(x)} = \frac{\left(\sum_j c_j\right) \texttt{Var}(x^*)}{\texttt{Var}(x)} = \frac{\left(\sum_j c_j\right)^2 \texttt{Var}(x^*)}{\left(\sum_j c_j\right) \texttt{Var}(x)} = \frac{\sigma}{\left(\sum_j c_j\right)} \\
\frac{\delta^*}{\delta} &= \frac{\texttt{Cov}(x^*, w)}{\texttt{Cov}(x, w)} = \frac{1}{\sum_j c_j}
\end{aligned}$$

where $\sigma := \left(\sum_j c_j\right)^2 \texttt{Var}(x^*)/\texttt{Var}(x)$ is the ratio of the signal variance to total variance. Substituting these into the formula for the plim of the OLS estimate for β:

$$\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} = \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2}\right] = \beta \left[\frac{\frac{\sigma}{\sum_j c_j} - \frac{1}{\sum_j c_j} R_{x|w}^2}{1 - R_{x|w}^2}\right] = \beta \left[\frac{1}{\sum_j c_j} \cdot \frac{\sigma - R_{x|w}^2}{1 - R_{x|w}^2}\right]$$

If $w = 0$, the regression is $y = \alpha + \beta x^* + \varepsilon$. The formula simplifies as $R_{x^*|w}^2 = R_{x|w}^2 = 0$.

$$\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} = \beta \left[\frac{\sigma}{\sum_j c_j}\right].$$

Substituting these into the formula for the plim of the OLS estimate for γ:

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2}\right] = \gamma + \beta\delta \left[\frac{\frac{1}{\sum_j c_j} - \frac{\sigma}{\sum_j c_j}}{1 - R_{x|w}^2}\right] = \gamma + \beta\delta \left[\frac{1}{\sum_j c_j} \cdot \frac{1 - \sigma}{1 - R_{x|w}^2}\right] = \gamma + \beta\delta^* \left[\frac{1 - \sigma}{1 - R_{x|w}^2}\right].$$

Non-standardized Manifest Variables. When PCA is applied using the covariance matrix of the z's, the derivation of the OLS plim is identical to above except $c_j = v_j$.

Formative-Indicators Model

Standardized and Uncorrelated Manifest Variables. Let $\mathtt{Cov}(z_j, z_k) = 0\ \forall j \neq k$. The latent variable is

$$x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^* z_j.$$

Let v be the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the standardized z's. The standardized z's are given by $Z_j = \frac{z_j - \mathtt{E}[z_j]}{\sigma_{z_j}}$. The index x is the first principal component, $x := \sum_j v_j Z_j$, where $v_j = 0$, $j = J+1, ..., \mathcal{J}$. The regression model is

$$y = \alpha + \beta x^* + \gamma w + \varepsilon.$$

From Section C.1 the plim of the OLS estimates replacing x^* with x is given by:

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} &= \beta\left[\frac{\delta_x^* - \frac{\delta^*}{\delta}R_{x|w}^2}{1 - R_{x|w}^2}\right] \\
\mathtt{plim}\,\widehat{\gamma}^{\mathrm{OLS}} &= \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2}\right].
\end{aligned}$$

Since $x := \sum_j v_j Z_j$, we can express x as

$$x = \sum\nolimits_j v_j\left(\frac{z_j}{\sigma_{z_j}}\right) - \sum\nolimits_j v_j\left(\frac{\mathtt{E}[z_j]}{\sigma_{z_j}}\right).$$

Let $c_j := v_j/\sigma_j$. Then

$$x = \sum\nolimits_j c_j z_j - \mathtt{E}[x^*]\sum\nolimits_j c_j$$

and

$$\mu = \sum\nolimits_j \left(c_j - \lambda_j^*\right) z_j - \mathtt{E}[x^*]\sum\nolimits_j c_j.$$

The variance terms are:

$$\begin{aligned}
\mathtt{Var}(x^*) &= \mathtt{Var}\left(\sum\nolimits_j \lambda_j^* z_j\right) = \sum\nolimits_j (\lambda_j^*)^2 \mathtt{Var}(z_j) = \sum\nolimits_j (\lambda_j^*)^2 \sigma_{z_j}^2 \\
\mathtt{Var}(x) &= \mathtt{Var}\left(\sum\nolimits_j c_j z_j - \psi\right) = \mathtt{Var}\left(\sum\nolimits_j c_j z_j\right) = \sum\nolimits_j c_j^2 \mathtt{Var}(z_j) = \sum\nolimits_j \left(\frac{v_j^2}{\sigma_{z_j}^2}\right)\sigma_{z_j}^2 = \sum\nolimits_j v_j^2
\end{aligned}$$

where $\psi := \mathtt{E}[x^*]\sum_j c_j$. Since v is an eigenvector of the correlation matrix, the sum of squares of its elements is equal to 1, so $\sum_j v_j^2 = 1$. Thus, $\mathtt{Var}(x) = 1$. The covariance terms are:

$$\begin{aligned}
\mathtt{Cov}(x^*, x) &= \mathtt{Cov}\left(\sum\nolimits_j \lambda_j^* z_j, \sum\nolimits_j c_j z_j\right) = \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2 \\
\mathtt{Cov}(x^*, w) &= \mathtt{Cov}\left(\sum\nolimits_j \lambda_j^* z_j, w\right) = \sum\nolimits_j \lambda_j^* \mathtt{Cov}(z_j, w) \\
\mathtt{Cov}(x, w) &= \mathtt{Cov}\left(\sum\nolimits_j c_j z_j, w\right) = \sum\nolimits_j c_j \mathtt{Cov}(z_j, w)
\end{aligned}$$

The regression coefficients:

$$\begin{aligned} \delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2 \\ \frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)} \end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2} \right] = \beta \left[\frac{\sum_j \lambda_j^* c_j \sigma_{z_j}^2 - \left(\frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)} \right) R_{x|w}^2}{1 - R_{x|w}^2} \right].$$

If $w = 0$, the regression is $y = \alpha + \beta x^* + \varepsilon$. The formula simplifies as $R_{x^*|w}^2 = R_{x|w}^2 = 0$.

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2.$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2} \right] = \gamma + \beta\delta \left[\frac{\frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)} - \sum_j \lambda_j^* c_j \sigma_{z_j}^2}{1 - R_{x|w}^2} \right].$$

Standardized and Correlated Manifest Variables. We relax the assumption $\mathtt{Cov}(z_j, z_{j'}) = 0$ and allow for arbitrary covariances. The variance terms are:

$$\begin{aligned} \mathtt{Var}(x^*) &= \mathtt{Var}\left(\sum\nolimits_j \lambda_j^* z_j\right) = \sum\nolimits_j \sum\nolimits_k \lambda_j^* \lambda_k^* \mathtt{Cov}(z_j, z_k) \\ \mathtt{Var}(x) &= \mathtt{Var}\left(\sum\nolimits_j c_j z_j - \psi\right) = \sum\nolimits_j \sum\nolimits_k c_j c_k \mathtt{Cov}(z_j, z_k) = \varphi \end{aligned}$$

where $\psi := \mathtt{E}[x^*] \sum_j c_j$ and φ is the largest eigenvalue of the correlation matrix. The covariance terms are:

$$\begin{aligned} \mathtt{Cov}(x^*, x) &= \mathtt{Cov}\left(\sum\nolimits_j \lambda_j^* z_j, \sum\nolimits_j c_j z_j\right) = \sum\nolimits_j \sum\nolimits_k \lambda_j^* c_k \mathtt{Cov}(z_j, z_k) \\ \mathtt{Cov}(x^*, w) &= \mathtt{Cov}\left(\sum\nolimits_j \lambda_j^* z_j, w\right) = \sum\nolimits_j \lambda_j^* \mathtt{Cov}(z_j, w) \\ \mathtt{Cov}(x, w) &= \mathtt{Cov}\left(\sum\nolimits_j c_j z_j, w\right) = \sum\nolimits_j c_j \mathtt{Cov}(z_j, w) \end{aligned}$$

The regression coefficients:

$$\begin{aligned}
\delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \frac{1}{\varphi} \sum\nolimits_j \sum\nolimits_k \lambda_j^* c_k \mathtt{Cov}(z_j, z_k) \\
\frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)}
\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2} \right] = \beta \left[\frac{\frac{1}{\varphi} \sum_j \sum_k \lambda_j^* c_k \mathtt{Cov}(z_j, z_k) - \left(\frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)} \right) R_{x|w}^2}{1 - R_{x|w}^2} \right].$$

If $w = 0$, the regression is $y = \alpha + \beta x^* + \varepsilon$. The formula simplifies as $R_{x^*|w}^2 = R_{x|w}^2 = 0$.

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \left[\frac{1}{\varphi} \sum\nolimits_j \sum\nolimits_k \lambda_j^* c_k \mathtt{Cov}(z_j, z_k) \right].$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2} \right] = \gamma + \beta\delta \left[\frac{\frac{\sum_j \lambda_j^* \mathtt{Cov}(z_j, w)}{\sum_j c_j \mathtt{Cov}(z_j, w)} - \frac{1}{\varphi} \sum_j \sum_k \lambda_j^* c_k \mathtt{Cov}(z_j, z_k)}{1 - R_{x|w}^2} \right].$$

Non-standardized Manifest Variables. When PCA is applied using the covariance matrix of the z's, the derivation of the OLS plim is identical to above except $c_j = v_j$. This holds for the case where the z's are orthogonal and when they are correlated.

E OLS Properties of Alternative Indices

E.1 OLS Bias Using Unit Weight Indices

The results are summarized in the following proposition.

Proposition 7. *In the reflective model, under Assumptions 1 and 3 and replacing x^* with $x^{\overline{z}}$ in Equation (1), the OLS estimates converge to*

$$\begin{aligned}\texttt{plim}\,\widehat{\beta^{\overline{z}}} &= \beta\rho_w \\ \texttt{plim}\,\widehat{\gamma^{\overline{z}}} &= \gamma + \beta\delta\left[1 - \rho_w\right]\end{aligned}$$

where $\rho_w \in [0,1]$ is the conditional reliability ratio. In the formative model, under Assumptions 2 and 3, assuming $\texttt{Cov}(z_j, z_{j'}) = 0$ for all $j \neq j'$, assuming all $\mathcal{J}$ factors are observed, and replacing x^* with $x^{\overline{z}}$ in Equation (1), the OLS estimate of β on $x^{\overline{\overline{z}}}$ converges to

$$\begin{aligned}\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} &= \beta\left[\frac{J\left(\frac{\sum_{j=1}^{J}\lambda_j^*\texttt{Var}(z_j)}{\sum_{j=1}^{J}\texttt{Var}(z_j)} - \frac{\sum_{j=1}^{\mathcal{J}}\lambda_j^*\texttt{Cov}(w,z_j)}{\sum_{j=1}^{J}\texttt{Cov}(w,z_j)}R^2_{x|w}\right)}{1 - R^2_{x|w}}\right] \\ \texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} &= \gamma + \beta\delta\left[\frac{J\left(\frac{\sum_{j=1}^{\mathcal{J}}\lambda_j^*\texttt{Cov}(w,z_j)}{\sum_{j=1}^{J}\texttt{Cov}(w,z_j)} - \frac{\sum_{j=1}^{J}\lambda_j^*\texttt{Var}(z_j)}{\sum_{j=1}^{J}\texttt{Var}(z_j)}\right)}{1 - R^2_{x|w}}\right].\end{aligned}$$

Reflective-Indicators Model

Let the observed variables be $z_j = x^* + u_j$, where x^* is a latent variable and u_j is classical measurement error with $\texttt{E}[u_j] = 0$ and $\texttt{Var}(u_j) = \sigma_j^2$. We assume the errors are uncorrelated with x^*, w, and with each other: $\texttt{Cov}(x^*, u_j) = \texttt{Cov}(\mathrm{w}, u_j) = 0\ \forall j$ and $\texttt{Cov}(u_j, u_k) = 0\ \forall j \neq k$. The proxy for x^* is the mean of the non-standardized z's: $x = \frac{1}{J}\sum_j z_j$. The error in this proxy is defined as $\mu = x - x^*$. Substituting the definition of z_j into the expression for x, we get:

$$x = \frac{1}{J}\sum\nolimits_j (x^* + u_j) = x^* + \frac{1}{J}\sum\nolimits_j u_j$$

The proxy error is then $\mu = \frac{1}{J}\sum_j u_j$. The variance term is:

$$\texttt{Var}(x) = \texttt{Var}\left(x^* + \frac{1}{J}\sum\nolimits_j u_j\right) = \texttt{Var}(x^*) + \texttt{Var}\left(\frac{1}{J}\sum\nolimits_j u_j\right) + 2\texttt{Cov}\left(x^*, \frac{1}{J}\sum\nolimits_j u_j\right)$$

Since $\texttt{Cov}(x^*, u_j) = 0$ and $\texttt{Cov}(u_j, u_k) = 0$ for $j \neq k$, the last term is zero and the variance of the sum of errors is the sum of their variances:

$$\texttt{Var}(x) = \texttt{Var}(x^*) + \frac{1}{J^2}\sum\nolimits_j \sigma_j^2$$

The covariance terms are:

$$\begin{aligned}
\texttt{Cov}(x^*,\mu) &= \texttt{Cov}\left(x^*,\frac{1}{J}\sum\nolimits_j u_j\right) = \frac{1}{J}\sum\nolimits_j \texttt{Cov}(x^*,u_j) = 0\\
\texttt{Cov}(w,\mu) &= \texttt{Cov}\left(w,\frac{1}{J}\sum\nolimits_j u_j\right) = \frac{1}{J}\sum\nolimits_j \texttt{Cov}(w,u_j) = 0\\
\texttt{Cov}(x,w) &= \texttt{Cov}(x^*,w) + \texttt{Cov}(\mu,w) = \texttt{Cov}(x^*,w)\\
\texttt{Cov}(x^*,x) &= \texttt{Var}(x^*)
\end{aligned}$$

The regression coefficients:

$$\begin{aligned}
\delta^*_x &= \frac{\texttt{Cov}(x^*,x)}{\texttt{Var}(x)} = \frac{\texttt{Var}(x^*)}{\texttt{Var}(x)}\\
\frac{\delta^*}{\delta} &= \frac{\texttt{Cov}(x^*,w)}{\texttt{Cov}(x,w)} = 1
\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\begin{aligned}
\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} &= \beta\left[\frac{\delta^*_x - \frac{\delta^*}{\delta}R^2_{x|w}}{1-R^2_{x|w}}\right] = \beta\left[\frac{\frac{\texttt{Var}(x^*)}{\texttt{Var}(x)} - R^2_{x|w}}{1-R^2_{x|w}}\right] = \beta\left[\frac{\rho - R^2_{x|w}}{1-R^2_{x|w}}\right] = \beta\left[\frac{\frac{\texttt{Var}(x^*)}{\texttt{Var}(x)} - \frac{\texttt{Cov}(x,w)^2}{\texttt{Var}(x)\texttt{Var}(w)}}{1-\frac{\texttt{Cov}(x,w)^2}{\texttt{Var}(x)\texttt{Var}(w)}}\right]\\
&= \beta\left[\frac{\texttt{Var}(x^*) - \frac{\texttt{Cov}(x^*,w)^2}{\texttt{Var}(w)}}{\texttt{Var}(x) - \frac{\texttt{Cov}(x,w)^2}{\texttt{Var}(w)}}\right] = \beta\left[\frac{\texttt{Var}(x^*|w)}{\texttt{Var}(x|w)}\right] = \beta\rho_w,
\end{aligned}$$

where $\rho,\rho_w \in [0,1]$ are the unconditional and conditional reliability ratios, respectively. If $w=0$, the regression is $y=\alpha+\beta x^*+\varepsilon$. The formula is unchanged except ρ is the unconditional reliability ratio. Substituting these into the formula for the plim of the OLS estimate of γ:

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta^*_x}{1-R^2_{x|w}}\right] = \gamma + \delta\left[\beta - \texttt{plim}\,\widehat{\beta}\right] = \gamma + \beta\delta\left[1-\rho_w\right].$$

Formative-Indicators Model

The manifest variables z are independent, $\texttt{Cov}(z_j,z_k)=0\ \forall j\neq k$. The latent variable is

$$x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \lambda^*_j z_j.$$

The proxy for x^* is the mean of the J observed non-standardized z's: $x=\frac{1}{J}\sum_{j=1}^{J} z_j$. The proxy error is

$$\mu = \sum\nolimits_{j=1}^{J}\left(\frac{1}{J}-\lambda^*_j\right)z_j - \sum\nolimits_{j=J+1}^{\mathcal{J}}\lambda^*_j z_j = \sum\nolimits_{j=1}^{\mathcal{J}}\left(v_j-\lambda^*_j\right)z_j,$$

where

$$v_j = \begin{cases} 1/J & \text{if } j = 1, ..., J \\ 0 & \text{otherwise} \end{cases}$$

The variance terms are:

$$\begin{aligned} \mathtt{Var}(x^*) &= \sum\nolimits_{j=1}^{\mathcal{J}} (\lambda_j^*)^2 \mathtt{Var}(z_j) \\ \mathtt{Var}(x) &= \sum\nolimits_{j=1}^{\mathcal{J}} v_j^2 \mathtt{Var}(z_j) = \sum\nolimits_{j=1}^{J} \frac{1}{J^2} \mathtt{Var}(z_j) \end{aligned}$$

The covariance terms are:

$$\begin{aligned} \mathtt{Cov}(x^*, \mu) &= \sum\nolimits_{j=1}^{\mathcal{J}} \left(v_j - \lambda_j^* \right) \lambda_j^* \mathtt{Var}(z_j) = \sum\nolimits_{j=1}^{\mathcal{J}} v_j \lambda_j^* \mathtt{Var}(z_j) - \mathtt{Var}(x^*) \\ &= \sum\nolimits_{j=1}^{J} v_j \lambda_j^* \mathtt{Var}(z_j) - \mathtt{Var}(x^*) = \frac{1}{J} \sum\nolimits_{j=1}^{J} \lambda_j^* \mathtt{Var}(z_j) - 1 \\ \mathtt{Cov}(x, \mu) &= \sum\nolimits_{j=1}^{J} \left(v_j - \lambda_j^* \right) v_j \mathtt{Var}(z_j) \\ \mathtt{Cov}(x^*, x) &= 1 + \mathtt{Cov}(x^*, \mu) = 1 + \frac{1}{J} \sum\nolimits_{j=1}^{J} \lambda_j^* \mathtt{Var}(z_j) - 1 = \frac{1}{J} \sum\nolimits_{j=1}^{J} \lambda_j^* \mathtt{Var}(z_j) \\ \mathtt{Cov}(w, \mu) &= \sum\nolimits_{j=1}^{\mathcal{J}} \left(v_j - \lambda_j^* \right) \mathtt{Cov}(w, z_j) = \frac{1}{J} \sum\nolimits_{j=1}^{J} \mathtt{Cov}(w, z_j) - \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^* \mathtt{Cov}(w, z_j) \\ \mathtt{Cov}(w, x^*) &= \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^* \mathtt{Cov}(w, z_j) \\ \mathtt{Cov}(w, x) &= \frac{1}{J} \sum\nolimits_{j=1}^{J} \mathtt{Cov}(w, z_j) \end{aligned}$$

The regression coefficients:

$$\begin{aligned} \delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \frac{\mathtt{Var}(x^*) + \mathtt{Cov}(x^*, \mu)}{\mathtt{Var}(x)} = J \frac{\sum_{j=1}^{J} \lambda_j^* \mathtt{Var}(z_j)}{\sum_{j=1}^{J} \mathtt{Var}(z_j)} \\ \frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = J \frac{\sum_{j=1}^{\mathcal{J}} \lambda_j^* \mathtt{Cov}(w, z_j)}{\sum_{j=1}^{J} \mathtt{Cov}(w, z_j)} \end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\begin{aligned} \mathtt{plim}\, \widehat{\beta}^{\mathtt{OLS}} &= \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2} \right] \\ &= \beta \left[\frac{J \left(\frac{\sum_{j=1}^{J} \lambda_j^* \mathtt{Var}(z_j)}{\sum_{j=1}^{J} \mathtt{Var}(z_j)} - \frac{\sum_{j=1}^{\mathcal{J}} \lambda_j^* \mathtt{Cov}(w, z_j)}{\sum_{j=1}^{J} \mathtt{Cov}(w, z_j)} R_{x|w}^2 \right)}{1 - R_{x|w}^2} \right] \end{aligned}$$

If $w = 0$, the regression is $y = \alpha + \beta x^* + \varepsilon$. The formula simplifies as $R^2_{x^*|w} = R^2_{x|w} = 0$.

$$\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} = \beta\delta^*_x = \beta\left[J\frac{\sum_{j=1}^{J}\lambda^*_j\mathtt{Var}(z_j)}{\sum_{j=1}^{J}\mathtt{Var}(z_j)}\right].$$

If $\mathtt{Cov}(z_j, z_k) \neq 0$, then the formulas are unchanged except the value of δ^*_x will now incorporate the covariances between the manifest variables. Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\,\widehat{\gamma}^{\mathrm{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta^*_x}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta\left[\frac{J\left(\frac{\sum_{j=1}^{\mathcal{J}}\lambda^*_j\mathtt{Cov}(w,z_j)}{\sum_{j=1}^{J}\mathtt{Cov}(w,z_j)} - \frac{\sum_{j=1}^{J}\lambda^*_j\mathtt{Var}(z_j)}{\sum_{j=1}^{J}\mathtt{Var}(z_j)}\right)}{1 - R^2_{x|w}}\right].$$

E.2 OLS Bias Using Mean z-Score Indices

The results are summarized in the following proposition.

Proposition 8. *In the reflective model, under Assumptions 1 and 3 and replacing x^* with $x^{\overline{\overline{z}}}$, the OLS estimates converge to*

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} &= \beta\left[\frac{1}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)}\cdot\frac{\sigma - R^2_{x|w}}{1 - R^2_{x|w}}\right] \\
\mathtt{plim}\,\widehat{\gamma}^{\overline{z}} &= \gamma + \beta\delta^*\left[\frac{1-\sigma}{1 - R^2_{x|w}}\right],
\end{aligned}$$

where $\sigma := \left[\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)\right]^2 \frac{\mathtt{Var}(x^*)}{\mathtt{Var}(x)}$ is the ratio of the signal variance to total variance. In the formative model, under Assumptions 2 and 3, assuming all $\mathcal{J}$ factors are observed, and replacing x^* with $x^{\overline{\overline{z}}}$, the OLS estimates converge to

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} &= \beta\left[\frac{\sum_{j=1}^{J}\lambda^*_j\sigma_{z_j} - J\frac{\sum_{j=1}^{\mathcal{J}}\lambda^*_j\mathtt{Cov}(w,z_j)}{\sum_{j=1}^{J}\frac{1}{\sigma_{z_j}}\mathtt{Cov}(w,z_j)}R^2_{x|w}}{1 - R^2_{x|w}}\right] \\
\mathtt{plim}\,\widehat{\gamma}^{\mathrm{OLS}} &= \gamma + \beta\delta\left[\frac{J\frac{\sum_{j=1}^{\mathcal{J}}\lambda^*_j\mathtt{Cov}(w,z_j)}{\sum_{j=1}^{J}\frac{1}{\sigma_{z_j}}\mathtt{Cov}(w,z_j)} - \sum_{j=1}^{J}\lambda^*_j\sigma_{z_j}}{1 - R^2_{x|w}}\right].
\end{aligned}$$

Reflective-Indicators Model

Let the observed variables be $z_j = x^* + u_j$, where x^* is a latent variable and u_j is classical measurement error with $\mathtt{E}[u_j] = 0$ and $\mathtt{Var}(u_j) = \sigma_j^2$. We assume the errors are uncorrelated with x^*, w, and with each other: $\mathtt{Cov}(x^*, u_j) = \mathtt{Cov}(w, u_j) = 0\ \forall j$ and $\mathtt{Cov}(u_j, u_k) = 0\ \forall j \neq k$. The proxy for x^* is the mean of the standardized z's, denoted by $Z_j = \frac{z_j - \mathtt{E}[z_j]}{\sigma_{z_j}}$ where $\sigma_{z_j} = \sqrt{\mathtt{Var}(z_j)}$. The index is:

$$x = \frac{1}{J}\sum\nolimits_j Z_{ij} = \frac{1}{J}\sum\nolimits_j \frac{x^* + u_j - \mathtt{E}[x^*]}{\sigma_{z_j}} = \frac{1}{J}\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)x^* + \frac{1}{J}\sum\nolimits_j \frac{u_j}{\sigma_{z_j}} - \frac{1}{J}\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)\mathtt{E}[x^*].$$

The proxy error is then

$$\mu = \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)x^* + \frac{1}{J}\sum\nolimits_j \frac{u_j}{\sigma_{z_j}} - \frac{1}{J}\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)\mathtt{E}[x^*].$$

The variance terms are:

$$\begin{aligned}
\mathtt{Var}(x) &= \mathtt{Var}\left(\left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)x^* + \frac{1}{J}\sum\nolimits_j \frac{u_j}{\sigma_{z_j}}\right) = \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)^2 \mathtt{Var}(x^*) + \frac{1}{J^2}\sum\nolimits_j \frac{\sigma_j^2}{\sigma_{z_j}^2} \\
\mathtt{Var}(\mu) &= \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)^2 \mathtt{Var}(x^*) + \frac{1}{J^2}\sum\nolimits_j \frac{\sigma_j^2}{\sigma_{z_j}^2}
\end{aligned}$$

The covariance terms are:

$$\begin{aligned}
\mathtt{Cov}(x^*, \mu) &= \mathtt{Cov}\left(x^*, \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)x^* + \frac{1}{J}\sum\nolimits_j \frac{u_j}{\sigma_{z_j}}\right) = \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)\mathtt{Var}(x^*) + \frac{1}{J}\sum\nolimits_j \frac{\mathtt{Cov}(x^*, u_j)}{\sigma_{z_j}} \\
&= \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)\mathtt{Var}(x^*) \\
\mathtt{Cov}(x^*, x) &= \frac{1}{J}\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)\mathtt{Var}(x^*) \\
\mathtt{Cov}(w, \mu) &= \mathtt{Cov}\left(w, \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)x^* + \frac{1}{J}\sum\nolimits_j \frac{u_j}{\sigma_{z_j}}\right) = \left(\frac{1}{J}\sum\nolimits_j \frac{1}{\sigma_{z_j}} - 1\right)\mathtt{Cov}(w, x^*)
\end{aligned}$$

The regression coefficients:

$$\begin{aligned}
\delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \frac{1}{J}\left(\sum\nolimits_j \frac{1}{\sigma_{z_j}}\right)\frac{\mathtt{Var}(x^*)}{\mathtt{Var}(x)} = \frac{\sigma}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)} \\
\frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\mathtt{Cov}(x^*, w)}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)\mathtt{Cov}(x^*, w)} = \frac{1}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)}
\end{aligned}$$

where $\sigma := \left[\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)\right]^2 \frac{\mathtt{Var}(x^*)}{\mathtt{Var}(x)}$ is the ratio of the signal variance to total variance. Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \left[\frac{\delta^*_x - \frac{\delta^*}{\delta} R^2_{x|w}}{1 - R^2_{x|w}}\right] = \beta \left[\frac{1}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)} \cdot \frac{\sigma - R^2_{x|w}}{1 - R^2_{x|w}}\right].$$

If $w = 0$, the regression is $y = \alpha + \beta x^* + \varepsilon$. The formula simplifies as $R^2_{x^*|w} = R^2_{x|w} = 0$.

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta \left[\frac{\sigma}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)}\right].$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta^*_x}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta \left[\frac{\frac{1}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)} - \frac{\sigma}{\frac{1}{J}\left(\sum_j \frac{1}{\sigma_{z_j}}\right)}}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta^* \left[\frac{1 - \sigma}{1 - R^2_{x|w}}\right].$$

Formative-Indicators Model

The manifest variables z are independent, $\mathtt{Cov}(z_j, z_k) = 0\ \forall j \neq k$. The latent variable is

$$x^* = \sum_{j=1}^{\mathcal{J}} \lambda^*_j z_j.$$

The proxy for x^* is the mean of the standardized z's, denoted by $Z_j = \frac{z_j - \mathtt{E}[z_j]}{\sigma_{z_j}}$ where $\sigma_{z_j} = \sqrt{\mathtt{Var}(z_j)}$. The index is:

$$x = \frac{1}{J}\sum_j Z_{ij} = \frac{1}{J}\sum_{j=1}^{J} \frac{z_j - \mathtt{E}[z_j]}{\sigma_{z_j}} = \frac{1}{J}\left(\sum_{j=1}^{J} \frac{1}{\sigma_{z_j}} z_j\right) - \frac{1}{J}\sum_{j=1}^{J} \frac{\mathtt{E}[z_j]}{\sigma_{z_j}}.$$

The proxy error is

$$\mu = \frac{1}{J}\left(\sum_{j=1}^{J} \frac{1}{\sigma_{z_j}} z_j\right) - \frac{1}{J}\sum_{j=1}^{J} \frac{\mathtt{E}[z_j]}{\sigma_{z_j}} - \sum_{j=1}^{\mathcal{J}} \lambda^*_j z_j = \sum_{j=1}^{\mathcal{J}} \left(v_j - \lambda^*_j\right) z_j - \frac{1}{J}\sum_{j=1}^{J} \frac{\mathtt{E}[z_j]}{\sigma_{z_j}},$$

where

$$v_j = \begin{cases} 1/(J\sigma_{z_j}) & \text{if } j = 1, ..., J \\ 0 & \text{otherwise} \end{cases}$$

The variance terms are:

$$\begin{aligned}\mathtt{Var}(x^{*}) &= \sum\nolimits_{j=1}^{\mathcal{J}} (\lambda_j^{*})^{2}\mathtt{Var}(z_j)\\ \mathtt{Var}(x) &= \sum\nolimits_{j=1}^{\mathcal{J}} v_j^{2}\mathtt{Var}(z_j) = \sum\nolimits_{j=1}^{J} \frac{1}{(J\sigma_{z_j})^{2}}\mathtt{Var}(z_j) = \frac{1}{J}\end{aligned}$$

The covariance terms are:

$$\begin{aligned}\mathtt{Cov}(x^{*},\mu) &= \sum\nolimits_{j=1}^{\mathcal{J}} \left(v_j - \lambda_j^{*}\right)\lambda_j^{*}\mathtt{Var}(z_j) = \sum\nolimits_{j=1}^{\mathcal{J}} v_j\lambda_j^{*}\mathtt{Var}(z_j) - \mathtt{Var}(x^{*}) = \sum\nolimits_{j=1}^{J} v_j\lambda_j^{*}\mathtt{Var}(z_j) - \mathtt{Var}(x^{*})\\ \mathtt{Cov}(x^{*},x) &= \frac{1}{J}\left(\sum\nolimits_{j=1}^{J} \frac{1}{\sigma_{z_j}}\lambda_j^{*}\mathtt{Var}(z_j)\right)\\ \mathtt{Cov}(w,\mu) &= \sum\nolimits_{j=1}^{\mathcal{J}} \left(v_j - \lambda_j^{*}\right)\mathtt{Cov}(w,z_j)\\ \mathtt{Cov}(w,x^{*}) &= \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^{*}\mathtt{Cov}(w,z_j)\\ \mathtt{Cov}(w,x) &= \frac{1}{J}\sum\nolimits_{j=1}^{J} \frac{1}{\sigma_{z_j}}\mathtt{Cov}(w,z_j)\end{aligned}$$

The regression coefficients:

$$\begin{aligned}\delta_x^{*} &= \frac{\mathtt{Cov}(x^{*},x)}{\mathtt{Var}(x)} = \sum\nolimits_{j=1}^{J} \frac{\lambda_j^{*}}{\sigma_{z_j}}\mathtt{Var}(z_j) = \sum\nolimits_{j=1}^{J} \lambda_j^{*}\sigma_{z_j}\\ \frac{\delta^{*}}{\delta} &= \frac{\mathtt{Cov}(x^{*},w)}{\mathtt{Cov}(x,w)} = J\frac{\sum_{j=1}^{\mathcal{J}} \lambda_j^{*}\mathtt{Cov}(w,z_j)}{\sum_{j=1}^{J} \frac{1}{\sigma_{z_j}}\mathtt{Cov}(w,z_j)}\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} = \beta\left[\frac{\delta_x^{*} - \frac{\delta^{*}}{\delta}R^{2}_{x|w}}{1 - R^{2}_{x|w}}\right] = \beta\left[\frac{\sum_{j=1}^{J} \lambda_j^{*}\sigma_{z_j} - J\frac{\sum_{j=1}^{\mathcal{J}} \lambda_j^{*}\mathtt{Cov}(w,z_j)}{\sum_{j=1}^{J} \frac{1}{\sigma_{z_j}}\mathtt{Cov}(w,z_j)}R^{2}_{x|w}}{1 - R^{2}_{x|w}}\right].$$

If $w = 0$, the regression is $y = \alpha + \beta x^{*} + \varepsilon$. The formula simplifies as $R^{2}_{x^{*}|w} = R^{2}_{x|w} = 0$:

$$\mathtt{plim}\,\widehat{\beta}^{\mathrm{OLS}} = \beta\sum\nolimits_{j=1}^{J} \lambda_j^{*}\sigma_{z_j}.$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2} \right] = \gamma + \beta\delta \left[\frac{J \frac{\sum_{j=1}^{\mathcal{J}} \lambda_j^* \mathsf{Cov}(w, z_j)}{\sum_{j=1}^{J} \frac{1}{\sigma_{z_j}} \mathsf{Cov}(w, z_j)} - \sum_{j=1}^{J} \lambda_j^* \sigma_{z_j}}{1 - R_{x|w}^2} \right].$$

F Partial Least Squares

F.1 Derivation of PLS Weights

Standardized Manifest Variables

Weights when $J = 2$. Let z_1 and z_2 be standardized random variables with correlation ρ_{12}. y is an outcome variable (not standardized). Goal is to obtain $x = w_1 z_1 + w_2 z_2$ using PLS under two different normalizations: (i) unit-norm of the weights is one and (ii) $\texttt{Var}(x) = 1$. PLS maximizes the squared covariance between the index x and the outcome y:

$$\max_{w_1, w_2} [\texttt{Cov}(w_1 z_1 + w_2 z_2, y)]^2 \qquad \text{subject to normalization constraint}$$

This is equivalent to solving the generalized Rayleigh quotient:

$$\max_{w_1, w_2} \frac{[\texttt{Cov}(w_1 z_1 + w_2 z_2, y)]^2}{\texttt{Var}(w_1 z_1 + w_2 z_2)}$$

and then applying the normalization constraint. Variance and covariance terms are:

$$\begin{aligned} [\texttt{Cov}(x, y)]^2 &= [w_1 \texttt{Cov}(z_1, y) + w_2 \texttt{Cov}(z_2, y)]^2 \\ &= [w_1 \sigma_{z_1 y} + w_2 \sigma_{z_2 y}]^2 \\ \texttt{Var}(x) &= \texttt{Var}(w_1 z_1 + w_2 z_2) \\ &= w_1^2 \texttt{Var}(z_1) + w_2^2 \texttt{Var}(z_2) + 2 w_1 w_2 \texttt{Cov}(z_1, z_2) \\ &= w_1^2 + w_2^2 + 2 w_1 w_2 \rho_{12} \end{aligned}$$

where $\sigma_{z_1 y} = \texttt{Cov}(z_1, y)$ and $\sigma_{z_2 y} = \texttt{Cov}(z_2, y)$.

CASE 1: Unit-Norm Normalization
Objective function:

$$\max_{w_1, w_2} (w_1 \sigma_{z_1 y} + w_2 \sigma_{z_2 y})^2 \qquad \text{subject to } w_1^2 + w_2^2 = 1$$

By the Cauchy–Schwarz inequality:

$$|w_1 \sigma_{z_1 y} + w_2 \sigma_{z_2 y}| \leq \sqrt{w_1^2 + w_2^2} \sqrt{\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2} = \sqrt{\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2}$$

The maximum is achieved when (w_1, w_2) is proportional to $(\sigma_{z_1 y}, \sigma_{z_2 y})$. Under the constraint $w_1^2 + w_2^2 = 1$:

$$w_1^{(1)} = \pm \frac{\sigma_{z_1 y}}{\sqrt{\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2}}, \quad w_2^{(1)} = \pm \frac{\sigma_{z_2 y}}{\sqrt{\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2}}$$

where the signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. Both weights are positive if $\sigma_{z_1 y}, \sigma_{z_2 y} > 0$.

CASE 2: Unit-Variance Normalization ($\texttt{Var}(x) = 1$)

Objective function:

$$\max_{w_1,w_2}(w_1\sigma_{z_1y}+w_2\sigma_{z_2y})^2 \qquad \text{subject to } w_1^2+w_2^2+2w_1w_2\rho_{12}=1$$

In matrix notation, this is:

$$\max_{\mathbf{w}}\frac{(\mathbf{w}'\mathbf{c})^2}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{z}$ and $\mathbf{c}=[\sigma_{z_1y}\quad\sigma_{z_2y}]'$ The solution to this generalized Rayleigh quotient is:

$$\mathbf{w}=k\boldsymbol{\Sigma}^{-1}\mathbf{c}$$

for some scalar k determined by the normalization constraint and

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}&=\frac{1}{1-\rho_{12}^2}\begin{pmatrix}1&-\rho_{12}\\-\rho_{12}&1\end{pmatrix}\\ \boldsymbol{\Sigma}^{-1}\mathbf{c}&=\frac{1}{1-\rho_{12}^2}\begin{pmatrix}1&-\rho_{12}\\-\rho_{12}&1\end{pmatrix}\begin{pmatrix}\sigma_{z_1y}\\\sigma_{z_2y}\end{pmatrix}\\ &=\frac{1}{1-\rho_{12}^2}\begin{pmatrix}\sigma_{z_1y}-\rho_{12}\sigma_{z_2y}\\\sigma_{z_2y}-\rho_{12}\sigma_{z_1y}\end{pmatrix}\end{aligned}$$

The unit-variance normalization requires: $\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}=1$. Since $\mathbf{w}=k\boldsymbol{\Sigma}^{-1}\mathbf{c}$:

$$k^2(\boldsymbol{\Sigma}^{-1}\mathbf{c})'\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}\mathbf{c})=k^2\mathbf{c}'\boldsymbol{\Sigma}^{-1}\mathbf{c}=1$$

where

$$\begin{aligned}\mathbf{c}'\boldsymbol{\Sigma}^{-1}\mathbf{c}&=\frac{1}{1-\rho_{12}^2}\begin{pmatrix}\sigma_{z_1y}&\sigma_{z_2y}\end{pmatrix}\begin{pmatrix}1&-\rho_{12}\\-\rho_{12}&1\end{pmatrix}\begin{pmatrix}\sigma_{z_1y}\\\sigma_{z_2y}\end{pmatrix}\\ &=\frac{1}{1-\rho_{12}^2}\begin{pmatrix}\sigma_{z_1y}&\sigma_{z_2y}\end{pmatrix}\begin{pmatrix}\sigma_{z_1y}-\rho_{12}\sigma_{z_2y}\\\sigma_{z_2y}-\rho_{12}\sigma_{z_1y}\end{pmatrix}\\ &=\frac{1}{1-\rho_{12}^2}[\sigma_{z_1y}(\sigma_{z_1y}-\rho_{12}\sigma_{z_2y})+\sigma_{z_2y}(\sigma_{z_2y}-\rho_{12}\sigma_{z_1y})]\\ &=\frac{1}{1-\rho_{12}^2}[\sigma_{z_1y}^2+\sigma_{z_2y}^2-2\sigma_{z_1y}\sigma_{z_2y}\rho_{12}]\\ &=\frac{\sigma_{z_1y}^2+\sigma_{z_2y}^2-2\sigma_{z_1y}\sigma_{z_2y}\rho_{12}}{1-\rho_{12}^2}\end{aligned}$$

Solving for k^2 yields:

$$k^2=\frac{1-\rho_{12}^2}{\sigma_{z_1y}^2+\sigma_{z_2y}^2-2\sigma_{z_1y}\sigma_{z_2y}\rho_{12}}$$

Therefore:

$$k=\pm\sqrt{\frac{1-\rho_{12}^2}{\sigma_{z_1y}^2+\sigma_{z_2y}^2-2\sigma_{z_1y}\sigma_{z_2y}\rho_{12}}}$$

The weights are:

$$\begin{aligned}
w_1^{(2)} &= k \cdot \frac{\sigma_{z_1 y} - \rho_{12}\sigma_{z_2 y}}{1-\rho_{12}^2} \\
&= \pm\sqrt{\frac{1-\rho_{12}^2}{\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2 - 2\sigma_{z_1 y}\sigma_{z_2 y}\rho_{12}}} \cdot \frac{\sigma_{z_1 y} - \rho_{12}\sigma_{z_2 y}}{1-\rho_{12}^2} \\
&= \pm\frac{\sigma_{z_1 y} - \rho_{12}\sigma_{z_2 y}}{\sqrt{(1-\rho_{12}^2)(\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2 - 2\sigma_{z_1 y}\sigma_{z_2 y}\rho_{12})}}
\end{aligned}$$

Similarly:

$$w_2^{(2)} = \pm\frac{\sigma_{z_2 y} - \rho_{12}\sigma_{z_1 y}}{\sqrt{(1-\rho_{12}^2)(\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2 - 2\sigma_{z_1 y}\sigma_{z_2 y}\rho_{12})}}$$

Altogether:

$$w_1^{(2)} = \pm\frac{\sigma_{z_1 y} - \rho_{12}\sigma_{z_2 y}}{\sqrt{(1-\rho_{12}^2)(\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2 - 2\sigma_{z_1 y}\sigma_{z_2 y}\rho_{12})}}, \quad w_2^{(2)} = \pm\frac{\sigma_{z_2 y} - \rho_{12}\sigma_{z_1 y}}{\sqrt{(1-\rho_{12}^2)(\sigma_{z_1 y}^2 + \sigma_{z_2 y}^2 - 2\sigma_{z_1 y}\sigma_{z_2 y}\rho_{12})}}$$

where the signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. Unlike the unit-norm case, individual weights need not be positive even if $\sigma_{z_j y} > 0 \; \forall j$, since $w_j^{(2)}$ depends on $\mathbf{\Sigma}^{-1}\mathbf{c}$ rather than $\mathbf{c}$ alone.

Weights when $J > 2$. Let $z_1, ..., z_J$ be standardized random variables with correlation matrix $\mathbf{\Sigma}$. y is an outcome variable (not standardized). Goal is to obtain $x = \sum_j w_j z_j$ using PLS under two different normalizations: (i) unit-norm of the weights is one and (ii) $\texttt{Var}(x) = 1$. PLS maximizes the squared covariance between the index x and the outcome y:

$$\max_{\mathbf{w}} [\texttt{Cov}(\mathbf{w}'\mathbf{z}, y)]^2 \qquad \text{subject to normalization constraint}$$

This is equivalent to solving the generalized Rayleigh quotient:

$$\max_{\mathbf{w}} \frac{[\texttt{Cov}(\mathbf{w}'\mathbf{z}, y)]^2}{\texttt{Var}(\mathbf{w}'\mathbf{z})}$$

and then applying the normalization constraint. Variance and covariance terms are:

$$\begin{aligned}
[\texttt{Cov}(x, y)]^2 &= \left[\sum\nolimits_j w_j \texttt{Cov}(z_j, y)\right]^2 \\
&= (\mathbf{w}'\mathbf{c})^2 \\
&= \left(\sum\nolimits_j w_j \sigma_{z_j y}\right)^2 \\
\texttt{Var}(x) &= \texttt{Var}\left(\sum\nolimits_j w_j z_j\right) \\
&= \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \\
&= \sum\nolimits_j \sum\nolimits_k w_j w_k \texttt{Cov}(z_j, z_k)
\end{aligned}$$

where $\sigma_{z_j y} = \mathsf{Cov}(z_j, y)$ and

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_J \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \sigma_{z_1 y} \\ \sigma_{z_2 y} \\ \vdots \\ \sigma_{z_J y} \end{pmatrix}.$$

CASE 1: Unit-Norm Normalization
Objective function:

$$\max_{\mathbf{w}} (\mathbf{w}'\mathbf{c})^2 \quad \text{subject to } \mathbf{w}'\mathbf{w} = 1$$

By the Cauchy–Schwarz inequality:

$$|\mathbf{w}'\mathbf{c}| \leq \|\mathbf{w}\|_2 \|\mathbf{c}\|_2 = \|\mathbf{c}\|_2$$

The maximum is achieved when **w** is proportional to **c**. Under the constraint $\mathbf{w}'\mathbf{w} = 1$:

$$w_j^{(1)} = \pm \frac{c_j}{\sqrt{\sum_k c_k^2}}, \qquad j = 1, ..., J$$

where the signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. All weights are positive if $\sigma_{z_j y} > 0 \, \forall j$.

CASE 2: Unit-Variance Normalization
Objective function:

$$\max_{\mathbf{w}} (\mathbf{w}'\mathbf{c})^2 \qquad \text{subject to } \mathbf{w}'\mathbf{\Sigma}\mathbf{w} = 1$$

In matrix notation, this is:

$$\max_{\mathbf{w}} \frac{(\mathbf{w}'\mathbf{c})^2}{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}}$$

where $\mathbf{\Sigma}$ is the covariance matrix of **z** and $\mathbf{c} = [\sigma_{z_1 y} \quad \sigma_{z_2 y}]'$ The solution to this generalized Rayleigh quotient is:

$$\mathbf{w} = k\mathbf{\Sigma}^{-1}\mathbf{c}$$

for some scalar k determined by the normalization constraint. The unit-variance normalization requires: $\mathbf{w}'\mathbf{\Sigma}\mathbf{w} = 1$. Since $\mathbf{w} = k\mathbf{\Sigma}^{-1}\mathbf{c}$:

$$k^2(\mathbf{\Sigma}^{-1}\mathbf{c})'\mathbf{\Sigma}(\mathbf{\Sigma}^{-1}\mathbf{c}) = k^2\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} = 1$$

From $k^2\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} = 1$:

$$k^2 = \frac{1}{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}$$

Therefore:

$$k = \pm \frac{1}{\sqrt{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}}$$

The weights are:

$$\mathbf{w}^{(2)} = \pm\frac{\boldsymbol{\Sigma}^{-1}\mathbf{c}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}^{-1}\mathbf{c}}}$$

or componentwise:

$$w_j^{(2)} = \pm\frac{[\boldsymbol{\Sigma}^{-1}\mathbf{c}]_j}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}^{-1}\mathbf{c}}}, \quad j = 1, ..., J$$

where $[\boldsymbol{\Sigma}^{-1}\mathbf{c}]_j$ denotes the $j^{\mathtt{th}}$ component of the vector $\boldsymbol{\Sigma}^{-1}\mathbf{c}$. The signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. Unlike the unit-norm case, individual weights need not be positive even if $\sigma_{z_j y} > 0\ \forall j$, since the $j^{\mathtt{th}}$ weight depends on $[\boldsymbol{\Sigma}^{-1}\mathbf{c}]_j$ rather than c_j alone.

Non-Standardized Manifest Variables

Weights when $J > 2$. Let $z_1, ..., z_J$ be non-standardized random variables with covariance matrix $\boldsymbol{\Sigma}$. y is an outcome variable (not standardized). Goal is to obtain $x = \sum_j w_j z_j$ using PLS under two different normalizations: (i) unit-norm of the weights is one and (ii) $\mathtt{Var}(x) = 1$. PLS maximizes the squared covariance between the index x and the outcome y:

$$\max_{\mathbf{w}} [\mathtt{Cov}(\mathbf{w}'\mathbf{z}, y)]^2 \quad \text{subject to normalization constraint}$$

This is equivalent to solving the generalized Rayleigh quotient:

$$\max_{\mathbf{w}} \frac{[\mathtt{Cov}(\mathbf{w}'\mathbf{z}, y)]^2}{\mathtt{Var}(\mathbf{w}'\mathbf{z})}$$

and then applying the normalization constraint. Variance and covariance terms are:

$$\begin{aligned}
[\mathtt{Cov}(x, y)]^2 &= \left[\sum\nolimits_j w_j \mathtt{Cov}(z_j, y)\right]^2 \\
&= (\mathbf{w}'\mathbf{c})^2 \\
&= \left(\sum\nolimits_j w_j \sigma_{z_j y}\right)^2 \\
\mathtt{Var}(x) &= \mathtt{Var}\left(\sum\nolimits_j w_j z_j\right) \\
&= \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\
&= \sum\nolimits_j \sum\nolimits_k w_j w_k \mathtt{Cov}(z_j, z_k)
\end{aligned}$$

where $\sigma_{z_j y} = \mathtt{Cov}(z_j, y)$ and

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_J \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \sigma_{z_1 y} \\ \sigma_{z_2 y} \\ \vdots \\ \sigma_{z_J y} \end{pmatrix}.$$

CASE 1: Unit-Norm Normalization

Objective function:

$$\max_{\mathbf{w}}(\mathbf{w}'\mathbf{c})^2 \quad \text{subject to } \mathbf{w}'\mathbf{w} = 1$$

By the Cauchy–Schwarz inequality:

$$|\mathbf{w}'\mathbf{c}| \leq \|\mathbf{w}\|_2\|\mathbf{c}\|_2 = \|\mathbf{c}\|_2$$

The maximum is achieved when **w** is proportional to **c**. Under the constraint $\mathbf{w}'\mathbf{w} = 1$:

$$w_j^{(1)} = \pm\frac{c_j}{\sqrt{\sum_k c_k^2}}, \qquad j = 1, ..., J$$

where the signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. All weights are positive if $\sigma_{z_j y} > 0 \; \forall j$.

CASE 2: Unit-Variance Normalization

Objective function:

$$\max_{\mathbf{w}}(\mathbf{w}'\mathbf{c})^2 \qquad \text{subject to } \mathbf{w}'\mathbf{\Sigma}\mathbf{w} = 1$$

In matrix notation, this is:

$$\max_{\mathbf{w}} \frac{(\mathbf{w}'\mathbf{c})^2}{\mathbf{w}'\mathbf{\Sigma}\mathbf{w}}$$

where $\mathbf{\Sigma}$ is the covariance matrix of **z** and $\mathbf{c} = [\sigma_{z_1 y} \quad \sigma_{z_2 y}]'$ The solution to this generalized Rayleigh quotient is:

$$\mathbf{w} = k\mathbf{\Sigma}^{-1}\mathbf{c}$$

for some scalar k determined by the normalization constraint. The unit-variance normalization requires: $\mathbf{w}'\mathbf{\Sigma}\mathbf{w} = 1$. Since $\mathbf{w} = k\mathbf{\Sigma}^{-1}\mathbf{c}$:

$$k^2(\mathbf{\Sigma}^{-1}\mathbf{c})'\mathbf{\Sigma}(\mathbf{\Sigma}^{-1}\mathbf{c}) = k^2\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} = 1$$

From $k^2\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} = 1$:

$$k^2 = \frac{1}{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}$$

Therefore:

$$k = \pm\frac{1}{\sqrt{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}}$$

The weights are:

$$\mathbf{w}^{(2)} = \pm\frac{\mathbf{\Sigma}^{-1}\mathbf{c}}{\sqrt{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}}$$

or componentwise:

$$w_j^{(2)} = \pm\frac{[\mathbf{\Sigma}^{-1}\mathbf{c}]_j}{\sqrt{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}}, \quad j = 1, ..., J$$

where $[\mathbf{\Sigma}^{-1}\mathbf{c}]_j$ denotes the j^{th} component of the vector $\mathbf{\Sigma}^{-1}\mathbf{c}$. The signs are chosen consistently (both + or both -) to maximize $(\mathbf{w}'\mathbf{c})^2$. Unlike the unit-norm case, individual weights need not be positive even if $\sigma_{z_j y} > 0\ \forall j$, since the j^{th} weight depends on $[\mathbf{\Sigma}^{-1}\mathbf{c}]_j$ rather than c_j alone. Note: The formulas do not change when the manifest variables are standardized.

F.2 OLS Properties: Reflective-Indicators Model

The results are summarized in the following proposition.

Proposition 9. *In the reflective model, under Assumptions 1 and 3 and replacing x^* with x^{PLS} derived under the unit variance normalization, the OLS estimates converge to*

$$\texttt{plim}\,\widehat{\beta}^{\text{OLS}} = \beta \left[\frac{1}{\sum_j c_j} \cdot \frac{\left(\sum_j c_j\right)^2 - R^2_{x|w}}{1 - R^2_{x|w}} \right]$$

and

$$\texttt{plim}\,\widehat{\gamma}^{\text{OLS}} = \gamma + \beta\delta \left[\frac{1}{\sum_j c_j} \cdot \frac{1 - \left(\sum_j c_j\right)^2}{1 - R^2_{x|w}} \right],$$

where v_j are the PLS weights and $c_j := v_j/\sigma_{z_j}$ if the manifest variables are standardized and $c_j := v_j$ if not.

F.2.1 Standardized Manifest Variables

Let the observed variables be $z_j = x^* + u_j$, where x^* is a latent variable and u_j is classical measurement error with $\texttt{E}[u_j] = 0$ and $\texttt{Var}(u_j) = \sigma_j^2$. We assume the errors are uncorrelated with x^*, w, and with each other: $\texttt{Cov}(x^*, u_j) = \texttt{Cov}(w, u_j) = 0\ \forall j$ and $\texttt{Cov}(u_j, u_k) = 0\ \forall j \neq k$. The proxy for x^* is obtained by applying PLS to the standardized manifest variables denoted by $Z_j = \frac{z_j - \texttt{E}[z_j]}{\sigma_{z_j}}$ where $\sigma_{z_j} = \sqrt{\texttt{Var}(z_j)}$. The error in this proxy is defined as $\mu = x - x^*$. Let $\sigma_j := \texttt{Cov}(Z_j, y)$. Let $\boldsymbol{\sigma}' = [\sigma_1 \ \cdots \ \sigma_J]$. PLS finds weights to maximize:

$$\max_{\mathbf{v}} \texttt{Cov}(\mathbf{v}'\mathbf{Z}, y) \quad \text{subject to} \quad \texttt{Var}(\mathbf{v}'\mathbf{Z}) = 1$$

under the unit variance normalization. The solution is:

$$v_j = \frac{[\mathbf{R}^{-1}\boldsymbol{\sigma}]_j}{\sqrt{\boldsymbol{\sigma}'\mathbf{R}^{-1}\boldsymbol{\sigma}}}, \quad j = 1, \ldots, J$$

where $\mathbf{R}$ is the correlation matrix of the z's. The PLS index is:

$$x = \sum_j v_j Z_j.$$

The proxy error is then:

$$\begin{aligned}\mu = x - x^* = \sum\nolimits_j v_j Z_j - x^* &= \sum\nolimits_j v_j \left(\frac{x^* + u_j - \mathtt{E}[x^*]}{\sigma_{z_j}}\right) - x^* \\ &= \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}}\right)(x^* - \mathtt{E}[x^*]) + \sum\nolimits_j \frac{v_j u_j}{\sigma_{z_j}} - x^* \\ &= \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} - 1\right) x^* + \sum\nolimits_j \left(\frac{v_j}{\sigma_{z_j}}\right) u_j - \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} - 1\right) \mathtt{E}[x^*]\end{aligned}$$

and

$$\begin{aligned}x &= \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}}\right) x^* + \sum\nolimits_j \left(\frac{v_j}{\sigma_{z_j}}\right) u_j - \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} - 1\right) \mathtt{E}[x^*] \\ &:= \sum\nolimits_j c_j x^* + \psi,\end{aligned}$$

where $c_j := v_j / \sigma_{z_j}$ and $\psi := \sum_j \left(\frac{v_j}{\sigma_{z_j}}\right) u_j - \left(\sum_j \frac{v_j}{\sigma_{z_j}} - 1\right) \mathtt{E}[x^*]$. The covariance terms are:

$$\begin{aligned}\mathtt{Cov}(x^*, x) &= \mathtt{Cov}\left(x^*, \sum\nolimits_j v_j Z_j\right) = \sum\nolimits_j v_j \mathtt{Cov}\left(x^*, \frac{x^* + u_j - \mathtt{E}[x^*]}{\sigma_{z_j}}\right) \\ &= \mathtt{Var}(x^*) \sum\nolimits_j c_j \\ \mathtt{Cov}(x^*, \mu) &= \mathtt{Cov}(x^*, x - x^*) = \mathtt{Cov}(x^*, x) - \mathtt{Var}(x^*) \\ &= \mathtt{Var}(x^*) \sum\nolimits_j c_j - \mathtt{Var}(x^*) = \mathtt{Var}(x^*) \left(\sum\nolimits_j c_j - 1\right) \\ \mathtt{Cov}(w, x) &= \sum\nolimits_j c_j \mathtt{Cov}(w, z_j) = \left(\sum\nolimits_j c_j\right) \mathtt{Cov}(w, x^*) \\ \mathtt{Cov}(w, \mu) &= \sum\nolimits_j \left(c_j - 1\right) \mathtt{Cov}(w, x^*)\end{aligned}$$

The regression coefficients:

$$\begin{aligned}\delta^*_x &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \frac{\mathtt{Var}(x^*) \sum_j c_j}{\mathtt{Var}(x)} = \sum\nolimits_j c_j \\ \frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\mathtt{Cov}(x^*, w)}{\left(\sum_j c_j\right) \mathtt{Cov}(w, x^*)} = \frac{1}{\sum_j c_j}\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\, \widehat{\beta}^{\mathrm{OLS}} = \beta \left[\frac{\delta^*_x - \frac{\delta^*}{\delta} R^2_{x|w}}{1 - R^2_{x|w}}\right] = \beta \left[\frac{\sum_j c_j - \left(\frac{1}{\sum_j c_j}\right) R^2_{x|w}}{1 - R^2_{x|w}}\right] = \beta \left[\frac{1}{\sum_j c_j} \cdot \frac{\left(\sum_j c_j\right)^2 - R^2_{x|w}}{1 - R^2_{x|w}}\right]$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta\left[\frac{\frac{1}{\sum_j c_j} - \sum_j c_j}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta\left[\frac{1}{\sum_j c_j} \cdot \frac{1 - \left(\sum_j c_j\right)^2}{1 - R^2_{x|w}}\right]$$

F.2.2 Non-Standardized Manifest Variables

When PLS is applied using the non-standardized z's, the derivation of the OLS plim is identical to above except $c_j = v_j$.

F.3 OLS Properties: Formative-Indicators Model

The results are summarized in the following proposition.

Proposition 10. *In the formative model, under Assumptions 2 and 3, assuming* $\texttt{Cov}(z_j, z_k) = 0$ *for all* $j \neq k$*,, assuming* $J \leq \mathcal{J}$ *manifest variables are observed, and replacing* x^* *with* x^{PLS} *derived under the unit variance normalization, the OLS estimates converge to*

$$\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} = \beta\left[\frac{\sum_j \lambda_j^* c_j \sigma^2_{z_j} - \left(\frac{\sum_j \lambda_j^* \texttt{Cov}(w,z_j)}{\sum_j c_j \texttt{Cov}(w,z_j)}\right) R^2_{x|w}}{1 - R^2_{x|w}}\right]$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\sum_j \lambda_j^* \texttt{Cov}(w,z_j)}{\sum_j c_j \texttt{Cov}(w,z_j)} - \sum_j \lambda_j^* c_j \sigma^2_{z_j}}{1 - R^2_{x|w}}\right],$$

where v_j *are the PLS weights and* $c_j := v_j/\sigma_{z_j}$ *if the manifest variables are standardized and* $c_j := v_j$ *if not.*

F.3.1 Standardized Manifest Variables

Assume the manifest variables z are independent, $\texttt{Cov}(z_j, z_k) = 0\ \forall j \neq k$. The latent variable is

$$x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^* z_j.$$

The PLS weights based on the J observed manifest variables are identical to above:

$$v_j = \frac{[\mathbf{R}^{-1}\mathbf{c}]_j}{\sqrt{\mathbf{c}'\mathbf{R}^{-1}\mathbf{c}}}, \quad j = 1, \ldots, J$$

where $\mathbf{R}$ is the $J \times J$ correlation matrix of the z's and $v_j = 0$, $j = J+1, ..., \mathcal{J}$. The PLS index is:

$$x = \sum\nolimits_{j=1}^{J} v_j Z_j.$$

The proxy error is:

$$\mu = x - x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} - \lambda_j^* \right) z_j,$$

and

$$\begin{aligned} x &= \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} \right) x^* + \sum\nolimits_j \left(\frac{v_j}{\sigma_{z_j}} \right) u_j - \left(\sum\nolimits_j \frac{v_j}{\sigma_{z_j}} - 1 \right) \mathtt{E}[x^*] \\ &:= \sum\nolimits_j c_j x^* + \psi, \end{aligned}$$

where $c_j := v_j / \sigma_{z_j}$, $v_j = 0 \; \forall j > J$, and $\psi := \sum_j \left(\frac{v_j}{\sigma_{z_j}} \right) u_j - \left(\sum_j \frac{v_j}{\sigma_{z_j}} - 1 \right) \mathtt{E}[x^*]$. The covariance terms are:

$$\begin{aligned} \mathtt{Cov}(x^*, x) &= \mathtt{Cov}\left(\sum\nolimits_j \lambda_j^* z_j, \sum\nolimits_j c_j z_j \right) = \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2 \\ \mathtt{Cov}(x^*, \mu) &= \mathtt{Cov}(x^*, x - x^*) = \mathtt{Cov}(x^*, x) - \mathtt{Var}(x^*) = \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2 - \mathtt{Var}(x^*) \\ \mathtt{Cov}(w, x) &= \mathtt{Cov}\left(w, \sum\nolimits_j c_j z_j \right) = \sum\nolimits_j c_j \mathtt{Cov}(w, z_j) \\ \mathtt{Cov}(w, x^*) &= \sum\nolimits_j \lambda_j^* \mathtt{Cov}(w, z_j) \\ \mathtt{Cov}(w, \mu) &= \mathtt{Cov}(w, x) - \mathtt{Cov}(w, x^*) = \sum\nolimits_j \left(c_j - \lambda_j^* \right) \mathtt{Cov}(w, z_j) \end{aligned}$$

The regression coefficients:

$$\begin{aligned} \delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \sum\nolimits_j \lambda_j^* c_j \sigma_{z_j}^2 \\ \frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\sum_j \lambda_j^* \mathtt{Cov}(w, z_j)}{\sum_j c_j \mathtt{Cov}(w, z_j)} \end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\, \widehat{\beta}^{\text{OLS}} = \beta \left[\frac{\delta_x^* - \frac{\delta^*}{\delta} R_{x|w}^2}{1 - R_{x|w}^2} \right] = \beta \left[\frac{\sum_j \lambda_j^* c_j \sigma_{z_j}^2 - \left(\frac{\sum_j \lambda_j^* \mathtt{Cov}(w, z_j)}{\sum_j c_j \mathtt{Cov}(w, z_j)} \right) R_{x|w}^2}{1 - R_{x|w}^2} \right]$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\mathtt{plim}\, \widehat{\gamma}^{\text{OLS}} = \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta_x^*}{1 - R_{x|w}^2} \right] = \gamma + \beta\delta \left[\frac{\frac{\sum_j \lambda_j^* \mathtt{Cov}(w, z_j)}{\sum_j c_j \mathtt{Cov}(w, z_j)} - \sum_j \lambda_j^* c_j \sigma_{z_j}^2}{1 - R_{x|w}^2} \right]$$

F.3.2 Non-Standardized Manifest Variables

When PLS is the non-standardized z's, the derivation of the OLS plim is identical to above except $c_j = v_j$.

G Exploratory Factor Analysis

Because EFA relies on the correlation matrix of the manifest variables, we restrict attention to the standardized case.

G.1 OLS Properties: Reflective-Indicators Model With Standardized Manifest Variables

The results are summarized in the following proposition.

Proposition 11. *In the reflective model, under Assumptions 1 and 3 and replacing x^* with $x^{\texttt{EFA}}$ derived using the principal factor method, the OLS estimates converge to*

$$\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} = \beta\left[\frac{1}{\rho}\cdot\frac{\rho^2 - R^2_{x|w}}{1 - R^2_{x|w}}\right]$$

and

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{1}{\rho}\cdot\frac{1-\rho^2}{1 - R^2_{x|w}}\right],$$

where $\rho := \sqrt{\boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda}}$, $\boldsymbol{\lambda}$ is the vector of factor loadings, and $\mathbf{R}$ is the correlation matrix of the manifest variables.

Let the observed variables be $z_j = x^* + u_j$, where x^* is a latent variable and u_j is classical measurement error with $\texttt{E}[u_j] = 0$ and $\texttt{Var}(u_j) = \sigma_j^2$. We assume the errors are uncorrelated with x^*, w, and with each other: $\texttt{Cov}(x^*, u_j) = \texttt{Cov}(w, u_j) = 0\ \forall j$ and $\texttt{Cov}(u_j, u_k) = 0\ \forall j \neq k$. The proxy for x^* is obtained by applying EFA using the principal factor to the standardized manifest variables denoted by $Z_j = \frac{z_j - \texttt{E}[z_j]}{\sigma_{z_j}}$ where $\sigma_{z_j} = \sqrt{\texttt{Var}(z_j)}$. The error in this proxy is defined as $\mu = x - x^*$. The correlation matrix of standardized variables is denoted by $\mathbf{R}$. The factor model for standardized variables is:

$$Z_j = \lambda_j f + e_j,$$

where f is the common factor and e_j is the unique factor. Given the correlation structure with one underlying factor, the loadings are:

$$\lambda_j = \sqrt{\frac{\texttt{Var}(x^*)}{\texttt{Var}(z_j)}}$$

which allows us to write:

$$\mathbf{Z} = \boldsymbol{\lambda}\frac{x^*}{\sqrt{\texttt{Var}(x^*)}} + \mathbf{e}.$$

The regression-based factor score weights are:

$$\omega = \mathbf{R}^{-1}\boldsymbol{\lambda}$$

This gives non-standardized factor scores:

$$\begin{aligned}
\widehat{f} = \omega'\mathbf{Z} &= (\mathbf{R}^{-1}\lambda)'\mathbf{Z} = \lambda'\mathbf{R}^{-1}\mathbf{Z} = \lambda'\mathbf{R}^{-1}\left(\lambda\frac{x^*}{\sqrt{\mathtt{Var}(x^*)}} + \mathbf{e}\right) \\
&= \lambda'\mathbf{R}^{-1}\lambda\left(\frac{x^*}{\sqrt{\mathtt{Var}(x^*)}}\right) + \psi
\end{aligned}$$

where $\psi := \lambda'\mathbf{R}^{-1}\mathbf{e}$. The non-standardized regression-based factor score $\widehat{f}$ has variance:

$$\mathtt{Var}(\widehat{f}) = \mathtt{Var}(\omega'\tilde{\mathbf{z}}) = \omega'\mathbf{R}\omega = (\mathbf{R}^{-1}\lambda)'\mathbf{R}(\mathbf{R}^{-1}\lambda) = \lambda'\mathbf{R}^{-1}\lambda$$

Standardizing the scores:

$$x = \frac{\widehat{f}}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} = \sqrt{\lambda'\mathbf{R}^{-1}\lambda}\left(\frac{x^*}{\sqrt{\mathtt{Var}(x^*)}}\right) + \frac{\psi}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} = \left(\frac{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}{\sqrt{\mathtt{Var}(x^*)}}\right)x^* + \frac{\psi}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}$$

such that $\mathtt{Var}(x) = 1$, where ψ is classical measurement error. Covariance terms are:

$$\begin{aligned}
\mathtt{Cov}(\widehat{f}, x^*) &= \omega'\mathtt{Cov}(\mathbf{Z}, x^*) = (\mathbf{R}^{-1}\lambda)'\lambda\sqrt{\mathtt{Var}(x^*)} = \lambda'\mathbf{R}^{-1}\lambda \cdot \sqrt{\mathtt{Var}(x^*)} \\
\mathtt{Cov}(x, x^*) &= \frac{\mathtt{Cov}(\widehat{f}, x^*)}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} = \sqrt{\lambda'\mathbf{R}^{-1}\lambda} \cdot \sqrt{\mathtt{Var}(x^*)} \\
\mathtt{Cov}(x^*, \mu) &= \mathtt{Cov}(x^*, x - x^*) = \mathtt{Cov}(x^*, x) - \mathtt{Var}(x^*) = \sqrt{\lambda'\mathbf{R}^{-1}\lambda} \cdot \sqrt{\mathtt{Var}(x^*)} - \mathtt{Var}(x^*) \\
\mathtt{Cov}(x, w) &= \left(\frac{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}{\sqrt{\mathtt{Var}(x^*)}}\right)\mathtt{Cov}(x^*, w) \\
\mathtt{Cov}(w, \mu) &= \left(\frac{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}{\sqrt{\mathtt{Var}(x^*)}} - 1\right) \cdot \mathtt{Cov}(w, x^*)
\end{aligned}$$

since $\mathtt{Cov}(Z_j, x^*) = \lambda_j\sqrt{\mathtt{Var}(x^*)}$. The regression coefficients:

$$\begin{aligned}
\delta_x^* &= \frac{\mathtt{Cov}(x^*, x)}{\mathtt{Var}(x)} = \sqrt{\lambda'\mathbf{R}^{-1}\lambda} := \rho \\
\frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \frac{\sqrt{\mathtt{Var}(x^*)}}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} := \frac{1}{\rho}
\end{aligned}$$

where $\rho := \sqrt{\lambda'\mathbf{R}^{-1}\lambda}$. Substituting these into the formula for the plim of the OLS estimate of β:

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} = \beta\left[\frac{\delta_x^* - \frac{\delta^*}{\delta}R^2_{x|w}}{1 - R^2_{x|w}}\right] = \beta\left[\frac{\rho - \frac{1}{\rho}R^2_{x|w}}{1 - R^2_{x|w}}\right] = \beta\left[\frac{1}{\rho} \cdot \frac{\rho^2 - R^2_{x|w}}{1 - R^2_{x|w}}\right]$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{\frac{\delta^*}{\delta} - \delta^*_x}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta\left[\frac{\frac{1}{\rho} - \rho}{1 - R^2_{x|w}}\right] = \gamma + \beta\delta\left[\frac{1}{\rho} \cdot \frac{1 - \rho^2}{1 - R^2_{x|w}}\right]$$

G.2 OLS Properties: Formative-Indicators Model With Standardized Manifest Variables

The results are summarized in the following proposition.

Proposition 12. *In the formative model, under Assumptions 2 and 3, assuming* $\texttt{Cov}(z_j, z_k) = 0$ *for all* $j \neq k$*, assuming* $J \leq \mathcal{J}$ *manifest variables are observed, and replacing* x^* *with* $x^{\texttt{EFA}}$ *in Equation (1), the OLS estimates converge to*

$$\texttt{plim}\,\widehat{\beta}^{\texttt{OLS}} = \beta\left[\frac{1}{\rho} \cdot \frac{\boldsymbol{\lambda}^{*\prime}\mathbf{D}^{1/2}\boldsymbol{\lambda} - \rho^2 \cdot \frac{\boldsymbol{\lambda}^{*\prime}\texttt{Cov}(\mathbf{z},w)}{\boldsymbol{\lambda}'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\texttt{Cov}(\mathbf{z},w)} \cdot R^2_{x|w}}{1 - R^2_{x|w}}\right]$$

Substituting these into the formula for the plim of the OLS estimate of γ*:*

$$\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} = \gamma + \beta\delta\left[\frac{1}{\rho} \cdot \frac{\rho^2 \cdot \frac{\boldsymbol{\lambda}^{*\prime}\texttt{Cov}(\mathbf{z},w)}{\boldsymbol{\lambda}'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\texttt{Cov}(\mathbf{z},w)} - \boldsymbol{\lambda}^{*\prime}\mathbf{D}^{1/2}\boldsymbol{\lambda}}{1 - R^2_{x|w}}\right],$$

where $\rho := \sqrt{\boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda}}$*,* $\boldsymbol{\lambda}$ *is the weight vector defined below,* $\mathbf{R}$ *is the correlation matrix of the manifest variables, and* $\mathbf{D}$ *is a diagonal matrix with* $\sigma^2_{z_j}$ *as the representative element.*

Remark. Under the independence assumption ($\texttt{Cov}(z_j, z_k) = 0$ for $j \neq k$), the population correlation matrix is $\mathbf{R} = \mathbf{I}$, which leaves no off-diagonal covariance for a conventional principal-factor extraction to recover. A literal application of EFA to such data would yield degenerate loadings. The proposition should therefore be read as characterizing the OLS plim when the index takes the regression-based factor-score form $x^{\texttt{EFA}} = \boldsymbol{\lambda}'\mathbf{R}^{-1}\mathbf{Z}/\sqrt{\boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda}}$ for a given weight vector $\boldsymbol{\lambda}$, regardless of whether $\boldsymbol{\lambda}$ is the literal output of a factor extraction. With $\mathbf{R} = \mathbf{I}$, this simplifies to a weighted average of standardized variables with weight proportional to λ_j. The reflective model parameterization $\lambda_j = \sqrt{\texttt{Var}(x^*)/\texttt{Var}(z_j)}$ is retained below to permit direct comparison with Proposition 11, but it should not be interpreted as the population EFA solution under the formative DGP.

The manifest variables z are independent, $\texttt{Cov}(z_j, z_k) = 0\ \forall j \neq k$. The latent variable is

$$x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \lambda^*_j z_j = \boldsymbol{\lambda}^{*\prime}\mathbf{z}.$$

Only $J \leq \mathcal{J}$ variables are observed. Using the parameterization from the reflective case, we set:

$$\lambda_j = \sqrt{\frac{\texttt{Var}(x^*)}{\texttt{Var}(z_j)}}.$$

The regression-based factor score weights are:

$$\omega = \mathbf{R}^{-1}\lambda$$

This gives non-standardized factor scores:

$$\widehat{f} = \omega'\mathbf{Z} \quad = \quad (\mathbf{R}^{-1}\lambda)'\mathbf{Z} = \lambda'\mathbf{R}^{-1}\mathbf{Z}.$$

The non-standardized regression-based factor score $\widehat{f}$ has variance:

$$\mathtt{Var}(\widehat{f}) = \mathtt{Var}(\omega'\tilde{\mathbf{z}}) = \omega'\mathbf{R}\omega = (\mathbf{R}^{-1}\lambda)'\mathbf{R}(\mathbf{R}^{-1}\lambda) = \lambda'\mathbf{R}^{-1}\lambda$$

Standardizing the scores:

$$x = \frac{\widehat{f}}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} = \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda'\mathbf{R}^{-1}\mathbf{Z}$$

such that $\mathtt{Var}(x) = 1$, where ψ is classical measurement error. Covariance terms are:

$$\begin{aligned}
\mathtt{Cov}(x^*, x) &= \mathtt{Cov}(\lambda^{*\prime}\mathbf{z},\ \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathbf{z}) = \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathtt{Cov}(\mathbf{z}, \mathbf{D}^{-1/2}\mathbf{z})\mathbf{R}^{-1}\lambda \\
&= \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathtt{Cov}(\mathbf{z}, \mathbf{z})\mathbf{D}^{-1/2}\mathbf{R}^{-1}\lambda = \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\boldsymbol{\Sigma}_{\mathbf{zz}}\mathbf{D}^{-1/2}\mathbf{R}^{-1}\lambda \\
&= \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{R}^{-1}\lambda = \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathbf{D}^{1/2}\lambda \\
\mathtt{Cov}(x^*, w) &= \lambda^{*\prime}\mathtt{Cov}(\mathbf{z}, w) \\
\mathtt{Cov}(x, w) &= \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathtt{Cov}(\mathbf{z}, w)
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with elements taken from $\boldsymbol{\Sigma}_{\mathbf{zz}}$. The regression coefficients:

$$\begin{aligned}
\delta^*_x &= \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathbf{D}^{1/2}\lambda \\
\frac{\delta^*}{\delta} &= \frac{\mathtt{Cov}(x^*, w)}{\mathtt{Cov}(x, w)} = \sqrt{\lambda'\mathbf{R}^{-1}\lambda} \cdot \frac{\lambda^{*\prime}\mathtt{Cov}(\mathbf{z}, w)}{\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathtt{Cov}(\mathbf{z}, w)}
\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of β:

$$\begin{aligned}
\mathtt{plim}\,\widehat{\beta}^{\mathtt{OLS}} &= \beta\left[\frac{\delta^*_x - \frac{\delta^*}{\delta}R^2_{x|w}}{1 - R^2_{x|w}}\right] \\
&= \beta\left[\frac{\frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathbf{D}^{1/2}\lambda - \sqrt{\lambda'\mathbf{R}^{-1}\lambda} \cdot \frac{\lambda^{*\prime}\mathtt{Cov}(\mathbf{z},w)}{\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathtt{Cov}(\mathbf{z},w)} \cdot R^2_{x|w}}{1 - R^2_{x|w}}\right] \\
&= \beta\left[\frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} \cdot \frac{\lambda^{*\prime}\mathbf{D}^{1/2}\lambda - \lambda'\mathbf{R}^{-1}\lambda \cdot \frac{\lambda^{*\prime}\mathtt{Cov}(\mathbf{z},w)}{\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathtt{Cov}(\mathbf{z},w)} \cdot R^2_{x|w}}{1 - R^2_{x|w}}\right]
\end{aligned}$$

Substituting these into the formula for the plim of the OLS estimate of γ:

$$\begin{aligned}
\texttt{plim}\,\widehat{\gamma}^{\texttt{OLS}} &= \gamma + \beta\delta \left[\frac{\frac{\delta^*}{\delta} - \delta^*_x}{1 - R^2_{x|w}} \right] = \gamma + \beta\delta \left[\frac{\sqrt{\lambda'\mathbf{R}^{-1}\lambda} \cdot \frac{\lambda^{*\prime}\mathsf{Cov}(\mathbf{z},w)}{\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathsf{Cov}(\mathbf{z},w)} - \frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}}\lambda^{*\prime}\mathbf{D}^{1/2}\lambda}{1 - R^2_{x|w}} \right] \\
&= \gamma + \beta\delta \left[\frac{1}{\sqrt{\lambda'\mathbf{R}^{-1}\lambda}} \cdot \frac{\lambda'\mathbf{R}^{-1}\lambda \cdot \frac{\lambda^{*\prime}\mathsf{Cov}(\mathbf{z},w)}{\lambda'\mathbf{R}^{-1}\mathbf{D}^{-1/2}\mathsf{Cov}(\mathbf{z},w)} - \lambda^{*\prime}\mathbf{D}^{1/2}\lambda}{1 - R^2_{x|w}} \right]
\end{aligned}$$

H Lubotsky–Wittenberg Approach

The plims are summarized in the following proposition.

Proposition 13. *In the reflective model, under Assumptions 1 and 3 except allowing* $\mathtt{Cov}(u_j, u_{j'}) \neq 0$ *for all* j, j'*, and replacing* x^* *with* $x = [z_1\ z_2\ \cdots z_J]$*, the OLS estimate* $\widehat{\beta}^{\mathtt{LW}}$ *converges to*

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{LW}} = \beta\left[\frac{\phi S_{uu}}{1+\phi S_{uu}}\right],$$

where $\phi := \mathtt{Var}(x^*|w)$ *and* $S_{uu} := \boldsymbol{\iota}'\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\iota}$ *with* $\boldsymbol{\iota}$ *being a* $J \times 1$ *vector of 1's is the sum of all elements of the inverse covariance matrix. The OLS estimate* $\widehat{\gamma}^{\mathtt{LW}}$ *converges to*

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{LW}} = \gamma + \beta\delta^*\left[\frac{1}{1+\phi S_{uu}}\right],$$

where δ^* *is the slope coefficient from the regression of* x^* *on* w*. In the formative model, under Assumptions 2 and 3, assuming* $J < \mathcal{J}$ *factors are observed, and replacing* x^* *with* $x = [z_1\ z_2\ \cdots z_J]$*, the OLS estimate* $\widehat{\beta}^{\mathtt{LW}}$ *converges to*

$$\mathtt{plim}\,\widehat{\beta}^{\mathtt{LW}} = \beta(\boldsymbol{\iota}'\boldsymbol{\lambda}^*) + \beta(\boldsymbol{\iota}'\boldsymbol{\Sigma}_{\mathbf{zz}\cdot w}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}v\cdot w}),$$

where $\boldsymbol{\iota}$ *is a* $J\times 1$ *vector of ones,* $\boldsymbol{\lambda}^*$ *is a* $J\times 1$ *vector of the weights on the observed z's, and* $v := \sum_{j=J+1}^{\mathcal{J}} \lambda_j^* z_j$*. The OLS estimate* $\widehat{\gamma}^{\mathtt{LW}}$ *converges to*

$$\mathtt{plim}\,\widehat{\gamma}^{\mathtt{LW}} = \gamma + \beta\Sigma_{ww\cdot\mathbf{z}}^{-1}\Sigma_{wv\cdot\mathbf{z}},$$

Proof: See Lubotsky and Wittenberg (2006) and Appendix H. ■

H.1 Reflective-Indicators Model

Let the observed variables be $z_j = x^* + u_j$, where x^* is a latent variable and u_j is classical measurement error with $\mathtt{E}[u_j] = 0$ and $\mathtt{Var}(u_j) = \sigma_j^2$. We assume the errors are uncorrelated with x^*, w, and with each other: $\mathtt{Cov}(x^*, u_j) = \mathtt{Cov}(w, u_j) = 0\ \forall j$ and $\mathtt{Cov}(u_j, u_k) = 0\ \forall j \neq k$. All variables are added to the regression model in lieu of x^*, with coefficient vector $\boldsymbol{\theta}$ and the estimate of β is $\boldsymbol{\iota}'\boldsymbol{\theta}$, where $\boldsymbol{\iota}$ is a column vector of ones. Covariance matrices are denoted by, for example, $\boldsymbol{\Sigma}_{\mathbf{zw}}$ and partial covariance matrices are denoted by, for example, $\boldsymbol{\Sigma}_{\mathbf{zz}\cdot\mathbf{w}}$. The true model is:

$$y = \alpha + \beta x^* + \gamma w + \varepsilon.$$

The estimated model is:

$$y = \theta_0 + \mathbf{z}'\boldsymbol{\theta} + w\gamma + u.$$

Let ϕ be the variance of x^* conditional on w:

$$\phi := \mathtt{Var}(x^*|w) = \mathtt{Var}(x^*) - \mathtt{Cov}(x^*, w)^2/\mathtt{Var}(w).$$

Let $\mathbf{\Sigma_{uu}}$ be the covariance matrix of the u's, which is diagonal. Let S_{uu} be the sum of the inverse variances:

$$S_{uu} := \sum_{j=1}^{J} \frac{1}{\sigma^2_{u_j}} = \boldsymbol{\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}.$$

The covariance of $\mathbf{z}$ conditional on w is

$$\mathbf{\Sigma}_{\mathbf{zz}\cdot w} = \mathbf{\Sigma_{uu}} + \phi\boldsymbol{\iota\iota}'$$

and the covariance between $\mathbf{z}$ and x^* conditional on w is:

$$\mathbf{\Sigma}_{\mathbf{z}x^*\cdot w} = \phi\boldsymbol{\iota}.$$

The inverse of $\mathbf{\Sigma}_{\mathbf{zz}\cdot w}$ equals $(A+uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}$ by the Sherman-Morrison formula. This yields:

$$\mathbf{\Sigma}^{-1}_{\mathbf{zz}\cdot w} = \mathbf{\Sigma}^{-1}_{\mathbf{uu}} - \frac{\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\phi}{1+\phi\boldsymbol{\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}} = \mathbf{\Sigma}^{-1}_{\mathbf{uu}} - \left(\frac{\phi}{1+\phi S_{uu}}\right)\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}$$

The plim for the coefficient vector is:

$$\begin{aligned}
\texttt{plim}\,\widehat{\boldsymbol{\theta}} &= \beta\mathbf{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\mathbf{\Sigma}_{\mathbf{z}x^*\cdot w} = \beta\mathbf{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\boldsymbol{\iota}\phi \\
&= \beta\phi\left[\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota} - \left(\frac{\phi}{1+\phi S_{uu}}\right)\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}(\boldsymbol{\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota})\right] \\
&= \beta\phi\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}\left[1 - \frac{\phi S_{uu}}{1+\phi S_{uu}}\right] = \beta\left[\frac{\phi}{1+\phi S_{uu}}\right]\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}.
\end{aligned}$$

Pre-multiplying by $\boldsymbol{\iota}'$ yields the coefficient sum:

$$\texttt{plim}\,\boldsymbol{\iota}'\widehat{\boldsymbol{\theta}} = \beta\left[\frac{\phi}{1+\phi S_{uu}}\right](\boldsymbol{\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}) = \beta\left[\frac{\phi S_{uu}}{1+\phi S_{uu}}\right].$$

If $w = 0$, ϕ becomes the unconditional variance of x^* which is normalized to one:

$$\texttt{plim}\,\boldsymbol{\iota}'\widehat{\boldsymbol{\theta}} = \beta\left[\frac{S_{uu}}{1+S_{uu}}\right].$$

If we allow $\texttt{Cov}(u_j, u_k) \neq 0\ \forall j, k$, then the plim is unchanged except now

$$S_{uu} := \boldsymbol{\iota}'\mathbf{\Sigma}^{-1}_{\mathbf{uu}}\boldsymbol{\iota}$$

which is the sum of all elements of the inverse covariance matrix.

To derive the plim for the estimate of γ, start with the OLS normal equation:

$$\texttt{plim}\,\widehat{\gamma} = \mathbf{\Sigma}^{-1}_{ww}\left(\mathbf{\Sigma}_{wy} - \mathbf{\Sigma}_{w\mathbf{z}}(\texttt{plim}\,\widehat{\boldsymbol{\theta}})\right),$$

where

$$
\begin{aligned}
\boldsymbol{\Sigma}_{wy} &= \mathtt{Cov}(w, \beta x^* + \gamma w + \varepsilon) = \beta \boldsymbol{\Sigma}_{wx^*} + \boldsymbol{\Sigma}_{ww}\gamma \\
\boldsymbol{\Sigma}_{w\mathbf{z}} &= \mathtt{Cov}(w, \boldsymbol{\iota} x^* + \mathbf{u}) = \boldsymbol{\Sigma}_{wx^*}\boldsymbol{\iota}' \\
\mathtt{plim}\,\widehat{\boldsymbol{\theta}} &= \beta \left[\frac{\phi}{1+\phi S_{uu}}\right] \boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\iota}.
\end{aligned}
$$

Substitution into the plim:

$$
\begin{aligned}
\mathtt{plim}\,\widehat{\gamma} &= \boldsymbol{\Sigma}_{ww}^{-1}\left[(\beta\boldsymbol{\Sigma}_{wx^*} + \boldsymbol{\Sigma}_{ww}\gamma) - (\boldsymbol{\Sigma}_{wx^*}\boldsymbol{\iota}')\left(\beta\frac{\phi}{1+\phi S_{uu}}\boldsymbol{\Sigma}_{\mathbf{uu}}^{-1}\boldsymbol{\iota}\right)\right] \\
&= \boldsymbol{\Sigma}_{ww}^{-1}\left[(\beta\boldsymbol{\Sigma}_{wx^*} + \boldsymbol{\Sigma}_{ww}\gamma) - \beta\boldsymbol{\Sigma}_{wx^*}\left(\frac{\phi S_{uu}}{1+\phi S_{uu}}\right)\right] \\
&= \gamma + \beta\left[1 - \frac{\phi S_{uu}}{1+\phi S_{uu}}\right]\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{wx^*} = \gamma + \beta\left[\frac{1}{1+\phi S_{uu}}\right](\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{wx^*}) \\
&= \gamma + \beta\delta^*\left[\frac{1}{1+\phi S_{uu}}\right].
\end{aligned}
$$

where δ^* is the slope coefficient from the regression of x^* on w.

H.2 Formative-Indicators Model

The latent variable is

$$x^* = \sum\nolimits_{j=1}^{\mathcal{J}} \lambda_j^* z_j.$$

Only $J \leq \mathcal{J}$ variables are observed, so we can write:

$$x^* = \sum\nolimits_{j=1}^{J} \lambda_j^* z_j + \sum\nolimits_{j=J+1}^{\mathcal{J}} \lambda_j^* z_j = \sum\nolimits_{j=1}^{J} \lambda_j^* z_j + \nu = \mathbf{z}'\boldsymbol{\lambda}^* + \nu,$$

where $\boldsymbol{\lambda}^*$ is a $J \times 1$ vector of the weights on the observed z's. The OLS plim of $\widehat{\boldsymbol{\theta}}$ is:

$$
\begin{aligned}
\mathtt{plim}\,\widehat{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\mathbf{zz}\cdot w}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}y\cdot w} \\
&= \boldsymbol{\Sigma}_{\mathbf{zz}\cdot w}^{-1}\beta\left(\boldsymbol{\Sigma}_{\mathbf{z}x^*} - \boldsymbol{\Sigma}_{\mathbf{z}w}\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{wx^*}\right) \\
&= \beta\boldsymbol{\Sigma}_{\mathbf{zz}\cdot w}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}x^*\cdot w}
\end{aligned}
$$

Covariance terms:

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{z}x^*} &= \mathtt{Cov}(\mathbf{z}, \mathbf{z}'\boldsymbol{\lambda}^* + \nu) = \boldsymbol{\Sigma}_{\mathbf{zz}}\boldsymbol{\lambda}^* + \boldsymbol{\Sigma}_{\mathbf{z}\nu} \\
\boldsymbol{\Sigma}_{wx^*} &= \mathtt{Cov}(w, \mathbf{z}'\boldsymbol{\lambda}^* + \nu) = \boldsymbol{\Sigma}_{w\mathbf{z}}\boldsymbol{\lambda}^* + \boldsymbol{\Sigma}_{w\nu} \\
\boldsymbol{\Sigma}_{\mathbf{z}x^*\cdot w} &= (\boldsymbol{\Sigma}_{\mathbf{zz}}\boldsymbol{\lambda}^* + \boldsymbol{\Sigma}_{\mathbf{z}\nu}) - \boldsymbol{\Sigma}_{\mathbf{z}w}\boldsymbol{\Sigma}_{ww}^{-1}(\boldsymbol{\Sigma}_{w\mathbf{z}}\boldsymbol{\lambda}^* + \boldsymbol{\Sigma}_{w\nu}) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{zz}} - \boldsymbol{\Sigma}_{\mathbf{z}w}\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{w\mathbf{z}}\right)\boldsymbol{\lambda}^* + \left(\boldsymbol{\Sigma}_{\mathbf{z}\nu} - \boldsymbol{\Sigma}_{\mathbf{z}w}\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{w\nu}\right)
\end{aligned}
$$

Substituting these into the formula for the plim:

$$\begin{aligned}\texttt{plim}\,\widehat{\boldsymbol{\theta}} &= \beta\boldsymbol{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\boldsymbol{\Sigma}_{\mathbf{z}x^{*}\cdot w}\\ &= \beta\boldsymbol{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\left(\boldsymbol{\Sigma}_{\mathbf{zz}\cdot w}\boldsymbol{\lambda}^{*} + \boldsymbol{\Sigma}_{\mathbf{z}v\cdot w}\right)\\ &= \beta\boldsymbol{\lambda}^{*} + \beta\boldsymbol{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\boldsymbol{\Sigma}_{\mathbf{z}v\cdot w}.\end{aligned}$$

The plim of the sum of the coefficients:

$$\texttt{plim}\,(\boldsymbol{\iota}'\widehat{\boldsymbol{\theta}}) = \beta(\boldsymbol{\iota}'\boldsymbol{\lambda}^{*}) + \beta(\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}_{\mathbf{zz}\cdot w}\boldsymbol{\Sigma}_{\mathbf{z}v\cdot w})$$

where the first term is the true coefficient sum and the second term reflects the omitted variable bias. If $w = 0$, the regression is $y = \alpha + \beta x^{*} + \varepsilon$. The partial covariances become unconditional covariances:

$$\texttt{plim}\,(\boldsymbol{\iota}'\widehat{\boldsymbol{\theta}}) = \beta(\boldsymbol{\iota}'\boldsymbol{\lambda}^{*}) + \beta(\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}_{\mathbf{zz}}\boldsymbol{\Sigma}_{\mathbf{z}v}).$$

The omitted variable bias term does not vanish simply because $w = 0$. If, in addition, $v = 0$ (i.e., $J = \mathcal{J}$, so all manifest variables are observed), it collapses to:

$$\texttt{plim}\,(\boldsymbol{\iota}'\widehat{\boldsymbol{\theta}}) = \beta(\boldsymbol{\iota}'\boldsymbol{\lambda}^{*}) = \beta\sum\nolimits_{j=1}^{J}\lambda^{*}_{j}.$$

The plim for the OLS estimate of γ is

$$\texttt{plim}\,\widehat{\gamma} \quad = \gamma + \beta\Sigma^{-1}_{ww\cdot\mathbf{z}}\Sigma_{wv\cdot\mathbf{z}},$$

which reflects the omitted variable bias from the unobserved z's.

I Hybrid IV Approach

Let the observed reflective indicators be $z_j = x^* + u_j$, where x^* is the latent variable and u_j is classical measurement error. We assume the errors are uncorrelated with x^* and w. The latent x^* is determined from the formative model, given by

$$x^* = \sum\nolimits_{\ell=1}^{\mathcal{L}} \lambda_\ell^* q_\ell,$$

where only $L \leq \mathcal{L}$ of the variables are observed. Two indices are created using some technique:

$$\begin{aligned} x &= \sum\nolimits_j v_j z_j \\ r &= \sum\nolimits_{\ell=1}^{L} \lambda_\ell q_\ell. \end{aligned}$$

Let $S_v := \sum_j v_j$. The true model is:

$$y = \alpha + \beta x^* + \gamma w + \varepsilon.$$

The estimated model is:

$$y = \alpha + \beta x + \gamma w + (\varepsilon - \beta\mu),$$

where

$$\begin{aligned} \mu := x - x^* &= \sum\nolimits_j v_j z_j - \sum\nolimits_{\ell=1}^{\mathcal{L}} \lambda_\ell^* q_\ell = \mathbf{z}'\mathbf{v} - \mathbf{q}'\boldsymbol{\lambda}^* \\ \nu := r - x^* &= \sum\nolimits_{\ell=1}^{\mathcal{L}} (\lambda_\ell - \lambda_\ell^*) q_\ell = \mathbf{q}'\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\right), \end{aligned}$$

where $\lambda_\ell = 0 \; \forall \ell > L$. The model is estimated using IV where r and w are the instruments. The plim of the IV estimate of β is:

$$\mathtt{plim}\, \widehat{\beta}^{\mathtt{IV}} = \frac{\mathtt{Cov}(r, y \mid w)}{\mathtt{Cov}(r, x \mid w)}$$

Covariance terms:

$$\begin{aligned} \mathtt{Cov}(r, y \mid w) &= \mathtt{Cov}(r, \beta x^* + \gamma w + \varepsilon \mid w) = \beta \mathtt{Cov}(r, x^* \mid w) \\ \mathtt{Cov}(r, x \mid w) &= \mathtt{Cov}(r, S_v x^* + \tilde{u} \mid w) = S_v \mathtt{Cov}(r, x^* \mid w) + \mathtt{Cov}(r, \tilde{u} \mid w) = S_v \mathtt{Cov}(r, x^* \mid w) \end{aligned}$$

Substitution yields:

$$\mathtt{plim}\, \widehat{\beta}^{\mathtt{IV}} = \frac{\beta \mathtt{Cov}(r, x^* \mid w)}{S_v \mathtt{Cov}(r, x^* \mid w)} = \frac{\beta}{S_v}$$

The IV estimate of γ is derived from the moment condition:

$$\mathtt{Cov}\left(w, y - \alpha - \widehat{\beta} x - \widehat{\gamma} w\right) = 0.$$

This implies

$$\widehat{\gamma} = \frac{\mathtt{Cov}(w, y)}{\mathtt{Var}(w)} - \widehat{\beta}\frac{\mathtt{Cov}(w, x)}{\mathtt{Var}(w)}$$

Covariance terms:

$$\begin{aligned}
\mathtt{Cov}(w, y) &= \mathtt{Cov}(w, \beta x^* + \gamma w + \varepsilon) = \beta\mathtt{Cov}(w, x^*) + \gamma\mathtt{Var}(w) \\
\mathtt{Cov}(w, x) &= \mathtt{Cov}\left(w, \sum\nolimits_j v_j z_j\right) = \mathtt{Cov}\left(w, \sum\nolimits_j v_j(x^* + u_j)\right) = S_v\mathtt{Cov}(w, x^*)
\end{aligned}$$

The plim of IV estimate of γ is:

$$\begin{aligned}
\mathtt{plim}\,\widehat{\gamma}^{\mathtt{IV}} &= \frac{\mathtt{Cov}(w, y)}{\mathtt{Var}(w)} - \left(\mathtt{plim}\,\widehat{\beta}^{\mathtt{IV}}\right)\frac{\mathtt{Cov}(w, x)}{\mathtt{Var}(w)} \\
&= \frac{\beta\mathtt{Cov}(w, x^*) + \gamma\mathtt{Var}(w)}{\mathtt{Var}(w)} - \frac{\beta}{S_v}\frac{\mathtt{Cov}(w, x)}{\mathtt{Var}(w)} \\
&= \gamma + \beta\left[\frac{\mathtt{Cov}(w, x^*)}{\mathtt{Var}(w)} - \frac{1}{S_v}\frac{S_v\mathtt{Cov}(w, x^*)}{\mathtt{Var}(w)}\right] = \gamma.
\end{aligned}$$

J Simulation Results

Table J.1

Scenario 1: Three Noisy Manifest Variables ($\beta = 0.25$)

Method	Bias (β)	RMSE (β)	Coverage (β)	Power (β)	Bias (η)	RMSE (η)	Coverage (η)	Power (η)
Panel A. N = 500								
True x^*	0.003	0.047	0.940	1.000	0.004	0.047	0.940	1.000
LW: Weighted	−0.058	0.071	0.660	0.996				
PLS Index (Std)	−0.037	0.058	0.874	0.994	−0.037	0.058	0.874	0.994
PLS Index (Non-Std)	−0.057	0.070	0.774	0.994	−0.057	0.070	0.774	0.994
YJL	−0.074	0.084	0.498	0.990	−0.030	0.055	0.926	0.990
PCA Index (Std)	−0.094	0.102	0.296	0.990	−0.060	0.077	0.736	0.990
IV ($z_1, z_2 \to z_3$)	0.005	0.078	0.946	0.988	0.594	0.646	0.224	0.988
LW	−0.074	0.084	0.506	0.988				
EFA Index	0.056	0.096	0.864	0.988	−0.056	0.074	0.770	0.988
IV ($z_1, z_3 \to z_2$)	0.003	0.065	0.954	0.986	0.368	0.400	0.308	0.986
YJL: Weights ∈ [0, 1]	−0.073	0.083	0.516	0.974	−0.031	0.056	0.910	0.974
Mean z-score Index	0.010	0.069	0.946	0.970	−0.068	0.083	0.678	0.970
Equal Weight Index	−0.158	0.161	0.002	0.908	−0.099	0.109	0.428	0.908
IV ($z_2, z_3 \to z_1$)	−0.001	0.099	0.948	0.754	0.054	0.132	0.928	0.754
PCA Index (Non-Std)	−0.222	0.222	0.000	0.542	−0.154	0.160	0.078	0.542
Panel B. N = 2000								
True x^*	0.000	0.021	0.958	1.000	0.000	0.022	0.958	1.000
IV ($z_1, z_3 \to z_2$)	0.001	0.031	0.954	1.000	0.363	0.370	0.002	1.000
Mean z-score Index	0.013	0.034	0.936	1.000	−0.066	0.070	0.170	1.000
IV ($z_1, z_2 \to z_3$)	0.001	0.035	0.946	1.000	0.581	0.593	0.000	1.000
PLS Index (Std)	−0.044	0.049	0.514	1.000	−0.044	0.049	0.514	1.000
EFA Index	0.055	0.066	0.642	1.000	−0.055	0.060	0.324	1.000
PLS Index (Non-Std)	−0.064	0.067	0.154	1.000	−0.064	0.067	0.154	1.000
LW: Weighted	−0.070	0.073	0.040	1.000				
YJL	−0.075	0.077	0.012	1.000	−0.038	0.044	0.638	1.000
LW	−0.075	0.078	0.016	1.000				
PCA Index (Std)	−0.094	0.096	0.000	1.000	−0.060	0.064	0.258	1.000
Equal Weight Index	−0.157	0.157	0.000	1.000	−0.096	0.099	0.008	1.000
IV ($z_2, z_3 \to z_1$)	0.004	0.049	0.946	0.998	0.061	0.085	0.814	0.998
YJL: Weights ∈ [0, 1]	−0.075	0.077	0.018	0.998	−0.038	0.044	0.628	0.998
PCA Index (Non-Std)	−0.220	0.221	0.000	0.986	−0.151	0.152	0.000	0.986

Notes.—The table presents the performance of various estimation methods for β in the model $y = \alpha + \beta x^* + \varepsilon$, where x^* is unobserved, and $\eta \coloneqq \beta \cdot \mathtt{SD}(\mathbf{x})$ where $\mathtt{SD}(\mathbf{x})$ is the empirical standard deviation of the index. Results based on 500 repetitions. Three manifest variables (z_1, z_2, and z_3) for x^* were generated with uncorrelated errors. The true value of β is 0.25, $\alpha = 0$, and η is determined from the empirical value of $\beta \cdot \mathtt{SD}(\mathbf{x}^*)$ in each data set. The table shows the average estimated β ($\widehat{\beta}$), bias, standard deviation of $\widehat{\beta}$ ($\sigma_{\widehat{\beta}}$), coverage probability of 95% confidence intervals (Coverage), and root mean squared error (RMSE) for each method; similarly for $\widehat{\eta}$. The IV rows report 2SLS estimates using two manifest variables as instruments for the third. LW and LW: Weighted include all three manifest variables and uses the sum or weighted sum of the three coefficient estimates. Values for η are missing for the LW estimators since there is no single index, x. YSL and YSL: Weights $\in [0, 1]$ include all three manifest variables and simultaneously estimates β and the weights using GMM. Values for η are obtained using the empirical standard deviation of the resulting index created using the estimated weights. Indices (PCA, PLS, Equal Weight, and Mean z-score) combine all manifest variables. LW = Lubotsky and Wittenberg (2006). YSL = Yang et al. (2023). Panels sorted by power and then RMSE for $\widehat{\beta}$.

Table J.2

Scenario 1: Three Noisy Manifest Variables ($\beta = 0$)

Method	Bias (β)	RMSE (β)	Coverage (β)	False Positive (β)	False Negative (β)	Bias (η)	RMSE (η)	Coverage (η)	False Positive (η)	False Negative (η)
Panel A. $N = 500$										
IV ($z_1, z_2 \rightarrow z_3$)	0.005	0.059	0.962	0.018	0.020	0.015	0.197	0.962	0.018	0.020
YJL	0.001	0.039	0.950	0.022	0.028	0.005	0.077	0.950	0.022	0.028
PCA Index (Non-Std)	0.010	0.013	0.950	0.050	0.000	0.035	0.044	0.950	0.050	0.000
IV ($z_2, z_3 \rightarrow z_1$)	−0.001	0.098	0.948	0.024	0.028	−0.002	0.119	0.948	0.024	0.028
IV ($z_1, z_3 \rightarrow z_2$)	0.003	0.058	0.948	0.026	0.026	0.008	0.143	0.948	0.026	0.026
LW	0.002	0.040	0.942	0.030	0.028					
Equal Weight Index	0.022	0.028	0.938	0.062	0.000	0.036	0.046	0.938	0.062	0.000
True x^*	0.003	0.047	0.936	0.032	0.032	0.003	0.047	0.936	0.032	0.032
EFA Index	0.059	0.075	0.934	0.066	0.000	0.037	0.047	0.934	0.066	0.000
Mean z-score Index	0.053	0.068	0.932	0.068	0.000	0.037	0.047	0.932	0.068	0.000
PCA Index (Std)	0.031	0.039	0.926	0.074	0.000	0.037	0.047	0.926	0.074	0.000
YJL: Weights $\in [0, 1]$	0.003	0.043	0.912	0.048	0.040	0.005	0.071	0.912	0.048	0.040
PLS Index (Non-Std)	0.067	0.073	0.750	0.250	0.000	0.067	0.073	0.750	0.250	0.000
PLS Index (Std)	0.070	0.076	0.748	0.252	0.000	0.070	0.076	0.748	0.252	0.000
LW: Weighted	−0.032	1.289	0.714	0.166	0.120					
Panel B. $N = 2000$										
True x^*	0.000	0.021	0.960	0.018	0.022	0.000	0.021	0.960	0.018	0.022
IV ($z_1, z_3 \rightarrow z_2$)	0.000	0.027	0.956	0.022	0.022	0.001	0.065	0.956	0.022	0.022
IV ($z_1, z_2 \rightarrow z_3$)	0.000	0.027	0.954	0.016	0.030	0.002	0.089	0.954	0.016	0.030
YJL	0.001	0.018	0.952	0.020	0.028	0.000	0.039	0.952	0.020	0.028
Mean z-score Index	0.025	0.032	0.950	0.050	0.000	0.018	0.022	0.950	0.050	0.000
PCA Index (Std)	0.014	0.018	0.950	0.050	0.000	0.018	0.022	0.950	0.050	0.000
LW	0.001	0.018	0.948	0.022	0.030					
PCA Index (Non-Std)	0.005	0.007	0.948	0.052	0.000	0.018	0.022	0.948	0.052	0.000
Equal Weight Index	0.010	0.013	0.948	0.052	0.000	0.017	0.022	0.948	0.052	0.000
EFA Index	0.028	0.035	0.946	0.054	0.000	0.018	0.022	0.946	0.054	0.000
IV ($z_2, z_3 \rightarrow z_1$)	0.004	0.047	0.942	0.030	0.028	0.005	0.058	0.942	0.030	0.028
YJL: Weights $\in [0, 1]$	0.000	0.021	0.920	0.032	0.048	0.000	0.035	0.920	0.032	0.048
PLS Index (Non-Std)	0.034	0.037	0.744	0.256	0.000	0.034	0.037	0.744	0.256	0.000
PLS Index (Std)	0.035	0.038	0.718	0.282	0.000	0.035	0.038	0.718	0.282	0.000
LW: Weighted	−0.001	0.524	0.694	0.128	0.178					

Notes.—See Table J.1. Panels sorted by coverage rate for $\widehat{\beta}$.

Table J.3

Scenario 2: Three Noisy Manifest Variables With An Additional Exogenous Covariate ($\beta = 0.25$, $\gamma = 0.25$)

Method	Bias (β)	RMSE (β)	Coverage (β)	Power (β)	Bias (η)	RMSE (η)	Coverage (η)	Power (η)	Bias (γ)	RMSE (γ)	Coverage (γ)	Power (γ)
Panel A. $N = 500$												
True x^*	0.001	0.050	0.966	0.996	0.001	0.051	0.958	0.998	0.003	0.051	0.958	0.998
IV $(z_1, z_3 \to z_2)$	0.000	0.073	0.970	0.968	0.361	0.402	0.548	0.968	0.005	0.063	0.950	0.948
LW: Weighted	−0.080	0.089	0.528	0.990					0.049	0.069	0.848	1.000
LW	−0.092	0.100	0.394	0.980					0.050	0.069	0.844	1.000
YJL: Weights $\in [0, 1]$	−0.090	0.098	0.428	0.964	−0.049	0.068	0.868	0.964	0.049	0.069	0.834	0.992
YJL	−0.092	0.100	0.394	0.988	−0.050	0.068	0.880	0.988	0.049	0.069	0.830	1.000
PLS Index (Std)	−0.059	0.076	0.780	0.982	−0.059	0.076	0.780	0.982	0.051	0.070	0.828	1.000
IV $(z_1, z_2 \to z_3)$	0.002	0.088	0.964	0.942	0.582	0.648	0.488	0.944	0.002	0.071	0.958	0.882
IV $(z_2, z_3 \to z_1)$	0.010	0.125	0.972	0.562	0.069	0.167	0.960	0.562	−0.001	0.078	0.962	0.864
EFA Index	0.020	0.077	0.956	0.944	−0.079	0.092	0.630	0.944	0.063	0.079	0.764	1.000
PCA Index (Std)	−0.113	0.120	0.210	0.934	−0.084	0.096	0.578	0.934	0.066	0.082	0.746	1.000
PLS Index (Non-Std)	−0.082	0.093	0.606	0.966	−0.082	0.093	0.606	0.966	0.069	0.083	0.730	1.000
Mean z-score Index	−0.023	0.071	0.956	0.906	−0.091	0.103	0.526	0.906	0.071	0.086	0.718	1.000
Equal Weight Index	−0.172	0.174	0.000	0.762	−0.122	0.130	0.300	0.762	0.090	0.102	0.542	1.000
PCA Index (Non-Std)	−0.227	0.227	0.000	0.388	−0.172	0.178	0.052	0.388	0.113	0.123	0.334	1.000
Panel B. $N = 2000$												
True x^*	0.001	0.026	0.956	1.000	0.002	0.026	0.966	1.000	−0.001	0.024	0.966	1.000
IV $(z_1, z_3 \to z_2)$	0.000	0.037	0.948	1.000	0.362	0.373	0.014	1.000	0.000	0.031	0.952	1.000
IV $(z_1, z_2 \to z_3)$	0.001	0.043	0.942	1.000	0.581	0.598	0.002	1.000	−0.001	0.034	0.964	1.000
IV $(z_2, z_3 \to z_1)$	0.002	0.063	0.956	0.990	0.059	0.098	0.882	0.990	−0.001	0.037	0.960	1.000
LW: Weighted	−0.088	0.091	0.014	1.000					0.045	0.051	0.576	1.000
YJL: Weights $\in [0, 1]$	−0.091	0.093	0.008	1.000	−0.056	0.062	0.402	1.000	0.045	0.051	0.584	1.000
YJL	−0.091	0.093	0.008	1.000	−0.057	0.062	0.392	1.000	0.045	0.051	0.574	1.000
LW	−0.091	0.093	0.010	1.000					0.045	0.051	0.566	1.000
PLS Index (Std)	−0.065	0.070	0.242	1.000	−0.065	0.070	0.242	1.000	0.049	0.055	0.498	1.000
EFA Index	0.019	0.044	0.926	1.000	−0.079	0.083	0.104	1.000	0.058	0.063	0.334	1.000
PCA Index (Std)	−0.113	0.115	0.000	1.000	−0.084	0.087	0.070	1.000	0.062	0.066	0.264	1.000
PLS Index (Non-Std)	−0.090	0.093	0.028	1.000	−0.090	0.093	0.028	1.000	0.067	0.071	0.188	1.000
Mean z-score Index	−0.024	0.043	0.884	1.000	−0.092	0.095	0.042	1.000	0.067	0.071	0.198	1.000
Equal Weight Index	−0.173	0.174	0.000	0.998	−0.124	0.126	0.000	0.998	0.087	0.090	0.034	1.000
PCA Index (Non-Std)	−0.227	0.227	0.000	0.908	−0.174	0.175	0.000	0.908	0.111	0.113	0.000	1.000

Notes.—See Table J.1 except the model is $y = \alpha + \beta x^* + \gamma w + \varepsilon$. Panels sorted by RMSE for $\widehat{\gamma}$.

TABLE J.4

SCENARIO 2: THREE NOISY MANIFEST VARIABLES WITH AN ADDITIONAL EXOGENOUS COVARIATE ($\beta = 0.25$, $\gamma = 0$)

Method	Bias (γ)	RMSE (γ)	Coverage (γ)	False Positive (γ)	False Negative (γ)
Panel A. $N = 500$					
True x^*	0.003	0.050	0.962	0.022	0.016
IV ($z_1, z_3 \to z_2$)	0.005	0.063	0.952	0.038	0.010
YJL	0.049	0.069	0.834	0.160	0.006
YJL: Weights $\in [0, 1]$	0.049	0.069	0.834	0.160	0.006
LW	0.049	0.069	0.844	0.150	0.006
LW: Weighted	0.050	0.070	0.840	0.154	0.006
IV ($z_1, z_2 \to z_3$)	0.003	0.070	0.960	0.032	0.008
PLS Index (Std)	0.052	0.071	0.822	0.174	0.004
IV ($z_2, z_3 \to z_1$)	0.000	0.076	0.968	0.024	0.008
EFA Index	0.062	0.078	0.770	0.230	0.000
PCA Index (Std)	0.065	0.080	0.756	0.244	0.000
PLS Index (Non-Std)	0.069	0.083	0.722	0.276	0.002
Mean z-score Index	0.070	0.085	0.730	0.270	0.000
Equal Weight Index	0.089	0.101	0.546	0.454	0.000
PCA Index (Non-Std)	0.113	0.122	0.336	0.664	0.000
Panel B. $N = 2000$					
True x^*	0.000	0.024	0.966	0.012	0.022
IV ($z_1, z_3 \to z_2$)	0.000	0.031	0.950	0.026	0.024
IV ($z_1, z_2 \to z_3$)	−0.002	0.034	0.966	0.024	0.010
IV ($z_2, z_3 \to z_1$)	−0.001	0.037	0.960	0.020	0.020
YJL	0.045	0.051	0.576	0.424	0.000
LW	0.045	0.051	0.574	0.426	0.000
LW: Weighted	0.045	0.051	0.566	0.434	0.000
YJL: Weights $\in [0, 1]$	0.045	0.051	0.574	0.426	0.000
PLS Index (Std)	0.050	0.055	0.490	0.510	0.000
EFA Index	0.059	0.063	0.328	0.672	0.000
PCA Index (Std)	0.062	0.067	0.266	0.734	0.000
PLS Index (Non-Std)	0.067	0.071	0.198	0.802	0.000
Mean z-score Index	0.067	0.071	0.194	0.806	0.000
Equal Weight Index	0.087	0.090	0.032	0.968	0.000
PCA Index (Non-Std)	0.110	0.113	0.000	1.000	0.000

NOTES.—See Table J.3.

Table J.5

Scenario 3: Three Noisy Manifest Variables On Different Scales With An Additional Exogenous Covariate ($\beta = 0.25$, $\gamma = 0.25$)

Method	Bias (β)	RMSE (β)	Coverage (β)	Power (β)	Bias (η)	RMSE (η)	Coverage (η)	Power (η)	Bias (γ)	RMSE (γ)	Coverage (γ)	Power (γ)
Panel A. $\omega_2 = 1.5$												
True x^*	−0.001	0.048	0.942	1.000	−0.001	0.048	0.950	1.000	0.003	0.047	0.950	1.000
LW: Weighted	−0.076	0.087	0.482	0.992					0.045	0.064	0.840	1.000
PLS Index (Std)	−0.057	0.072	0.770	0.992	−0.057	0.072	0.770	0.992	0.048	0.066	0.826	1.000
IV ($z_1, z_3 \rightarrow z_2$)	−0.084	0.095	0.480	0.984	0.196	0.231	0.626	0.984	0.003	0.056	0.954	0.990
PLS Index (Non-Std)	−0.074	0.085	0.630	0.982	−0.074	0.085	0.630	0.982	0.061	0.075	0.730	1.000
YJL	−0.096	0.104	0.286	0.980	−0.048	0.067	0.864	0.980	0.047	0.065	0.832	1.000
LW	−0.096	0.104	0.280	0.978					0.047	0.065	0.834	1.000
PCA Index (Std)	−0.103	0.109	0.206	0.978	−0.069	0.083	0.664	0.978	0.056	0.072	0.772	1.000
YJL: Weights ∈ [0, 1]	−0.095	0.103	0.290	0.978	−0.048	0.066	0.868	0.978	0.046	0.064	0.844	1.000
EFA Index	0.017	0.069	0.638	0.974	−0.067	0.081	0.676	0.974	0.055	0.070	0.782	1.000
Mean z-score Index	−0.016	0.066	0.946	0.962	−0.088	0.098	0.462	0.962	0.068	0.081	0.650	1.000
Equal Weight Index	−0.171	0.173	0.000	0.846	−0.120	0.128	0.216	0.846	0.088	0.098	0.462	1.000
IV ($z_2, z_3 \rightarrow z_1$)	0.008	0.093	0.952	0.806	0.066	0.131	0.918	0.808	−0.002	0.062	0.952	0.956
IV ($z_1, z_2 \rightarrow z_3$)	0.233	0.340	1.000	0.686	1.291	1.511	0.630	0.688	0.011	0.101	0.928	0.762
PCA Index (Non-Std)	−0.231	0.231	0.000	0.304	−0.187	0.192	0.014	0.304	0.118	0.125	0.208	1.000
Panel B. $\omega_2 = 5$												
True x^*	−0.002	0.048	0.946	1.000	−0.003	0.048	0.952	1.000	0.000	0.048	0.952	1.000
IV ($z_2, z_3 \rightarrow z_1$)	−0.003	0.054	0.958	1.000	0.052	0.083	0.886	1.000	0.001	0.049	0.956	1.000
PLS Index (Std)	−0.028	0.054	0.912	1.000	−0.028	0.054	0.912	1.000	0.020	0.050	0.932	1.000
PLS Index (Non-Std)	−0.028	0.054	0.906	1.000	−0.028	0.054	0.906	1.000	0.020	0.051	0.926	1.000
LW: Weighted	−0.031	0.057	0.872	1.000					0.019	0.050	0.938	1.000
PCA Index (Std)	−0.090	0.097	0.268	0.998	−0.036	0.059	0.878	0.998	0.026	0.053	0.922	1.000
EFA Index	0.002	0.054	0.902	0.998	−0.032	0.057	0.902	0.998	0.022	0.052	0.926	1.000
PCA Index (Non-Std)	−0.211	0.212	0.000	0.998	−0.035	0.058	0.886	0.998	0.024	0.053	0.918	1.000
Equal Weight Index	−0.173	0.174	0.000	0.996	−0.056	0.071	0.764	0.996	0.040	0.061	0.856	1.000
Mean z-score Index	0.003	0.059	0.946	0.988	−0.061	0.075	0.730	0.988	0.044	0.063	0.838	1.000
IV ($z_1, z_3 \rightarrow z_2$)	−0.200	0.201	0.000	0.968	0.021	0.071	0.940	0.968	−0.002	0.052	0.944	0.996
IV ($z_1, z_2 \rightarrow z_3$)	0.226	0.309	0.996	0.846	1.273	1.435	0.382	0.848	−0.002	0.097	0.954	0.756
YJL: Weights ∈ [0, 1]	−0.155	0.160	0.068	0.576	−0.020	0.050	0.994	0.576	0.019	0.051	0.920	1.000
LW	−0.158	0.164	0.068	0.542					0.020	0.051	0.930	1.000
YJL	−0.158	0.164	0.070	0.540	−0.024	0.062	1.000	0.540	0.018	0.051	0.916	1.000
Panel C. $\omega_2 = 10$												
True x^*	−0.001	0.048	0.948	0.998	−0.001	0.048	0.948	1.000	0.000	0.046	0.948	1.000
PLS Index (Non-Std)	−0.009	0.049	0.938	0.998	−0.009	0.049	0.938	0.998	0.007	0.046	0.940	1.000
PCA Index (Std)	−0.085	0.092	0.316	0.998	−0.026	0.054	0.910	0.998	0.021	0.050	0.926	1.000
LW: Weighted	−0.006	0.053	0.922	0.998					0.009	0.047	0.944	1.000
IV ($z_2, z_3 \rightarrow z_1$)	−0.002	0.051	0.948	0.998	0.054	0.082	0.844	0.998	0.001	0.047	0.944	1.000
EFA Index	0.004	0.053	0.936	0.998	−0.022	0.053	0.914	0.998	0.018	0.048	0.940	1.000
PLS Index (Std)	−0.018	0.050	0.934	0.998	−0.018	0.050	0.934	0.998	0.016	0.047	0.934	1.000
PCA Index (Non-Std)	−0.227	0.227	0.000	0.998	−0.011	0.051	0.932	0.998	0.009	0.047	0.946	1.000
Equal Weight Index	−0.194	0.194	0.000	0.996	−0.022	0.054	0.920	0.996	0.019	0.049	0.940	1.000
Mean z-score Index	0.009	0.063	0.952	0.986	−0.054	0.072	0.762	0.986	0.042	0.061	0.870	1.000
IV ($z_1, z_3 \rightarrow z_2$)	−0.225	0.225	0.000	0.978	0.004	0.062	0.950	0.978	0.002	0.048	0.958	1.000
IV ($z_1, z_2 \rightarrow z_3$)	0.250	0.353	0.998	0.858	1.347	1.562	0.304	0.858	0.002	0.092	0.944	0.734
YJL: Weights ∈ [0, 1]	−0.192	0.196	0.030	0.310	−0.005	0.048	0.998	0.310	0.007	0.046	0.954	0.994
YJL	−0.201	0.206	0.024	0.134	−0.020	0.070	1.000	0.134	0.000	0.051	0.940	1.000
LW	−0.207	0.214	0.026	0.124					0.008	0.047	0.946	1.000

Notes.—See Table J.3 except now $z_1 = x^* + u_1$, $z_2 = \omega_2 x^* + u_2$, and $z_3 = 0.5x^* + u_3$. Panels sorted by power of $\widehat{\beta}$.

Table J.6

Scenario 3: Three Noisy Manifest Variables On Different Scales With An Additional Exogenous Covariate
($\beta = 0.25$, $\gamma = 0$)

Method	Bias (γ)	RMSE (γ)	Coverage (γ)	False Positive (γ)	False Negative (γ)
Panel A. $\omega_2 = 1.5$					
True x^*	0.003	0.046	0.950	0.028	0.022
IV ($z_1, z_3 \rightarrow z_2$)	0.004	0.055	0.958	0.028	0.014
IV ($z_2, z_3 \rightarrow z_1$)	−0.002	0.062	0.948	0.024	0.028
LW: Weighted	0.046	0.064	0.842	0.158	0.000
YJL: Weights $\in [0, 1]$	0.045	0.064	0.840	0.160	0.000
PLS Index (Std)	0.048	0.065	0.832	0.168	0.000
YJL	0.046	0.065	0.830	0.170	0.000
LW	0.047	0.065	0.834	0.166	0.000
EFA Index	0.054	0.070	0.786	0.214	0.000
PCA Index (Std)	0.056	0.071	0.774	0.226	0.000
PLS Index (Non-Std)	0.062	0.076	0.720	0.280	0.000
Mean z-score Index	0.068	0.081	0.642	0.358	0.000
Equal Weight Index	0.088	0.098	0.468	0.532	0.000
IV ($z_1, z_2 \rightarrow z_3$)	0.009	0.101	0.930	0.070	0.000
PCA Index (Non-Std)	0.118	0.126	0.200	0.800	0.000
Panel B. $\omega_2 = 5$					
True x^*	−0.001	0.048	0.950	0.028	0.022
IV ($z_2, z_3 \rightarrow z_1$)	0.000	0.049	0.956	0.026	0.018
LW: Weighted	0.019	0.050	0.936	0.058	0.006
PLS Index (Non-Std)	0.020	0.050	0.932	0.062	0.006
YJL: Weights $\in [0, 1]$	0.019	0.050	0.924	0.068	0.008
LW	0.019	0.050	0.936	0.058	0.006
PLS Index (Std)	0.021	0.051	0.926	0.068	0.006
EFA Index	0.022	0.051	0.924	0.072	0.004
YJL	0.020	0.051	0.918	0.072	0.010
PCA Index (Std)	0.025	0.052	0.924	0.074	0.002
IV ($z_1, z_3 \rightarrow z_2$)	−0.001	0.052	0.944	0.028	0.028
PCA Index (Non-Std)	0.025	0.053	0.916	0.080	0.004
Equal Weight Index	0.039	0.060	0.858	0.142	0.000
Mean z-score Index	0.043	0.062	0.842	0.158	0.000
IV ($z_1, z_2 \rightarrow z_3$)	0.001	0.095	0.954	0.046	0.000
Panel C. $\omega_2 = 10$					
True x^*	0.001	0.045	0.952	0.024	0.024
YJL	0.008	0.045	0.958	0.030	0.012
LW: Weighted	0.008	0.046	0.948	0.032	0.020
PCA Index (Non-Std)	0.008	0.046	0.948	0.034	0.018
PLS Index (Non-Std)	0.007	0.046	0.940	0.036	0.024
YJL: Weights $\in [0, 1]$	0.007	0.046	0.956	0.030	0.014
IV ($z_2, z_3 \rightarrow z_1$)	0.001	0.047	0.944	0.028	0.028
PLS Index (Std)	0.015	0.047	0.940	0.050	0.010
LW	0.009	0.047	0.944	0.036	0.020
EFA Index	0.017	0.048	0.938	0.054	0.008
Equal Weight Index	0.018	0.048	0.946	0.050	0.004
PCA Index (Std)	0.021	0.049	0.934	0.060	0.006
IV ($z_1, z_3 \rightarrow z_2$)	0.001	0.049	0.956	0.024	0.020
Mean z-score Index	0.041	0.061	0.872	0.128	0.000
IV ($z_1, z_2 \rightarrow z_3$)	0.004	0.094	0.942	0.058	0.000

Notes.—See Table J.5. Panels sorted by RMSE of $\widehat{\gamma}$.

Table J.7

Scenario 5: Formative-Indicators Model With An Additional Exogenous Covariate (β = 0.25, γ = 0.25)

Method	Bias (β)	RMSE (β)	Coverage (β)	Power (β)	Bias (η)	RMSE (η)	Coverage (η)	Power (η)	Bias (γ)	RMSE (γ)	Coverage (γ)	Power (γ)
Panel A. $\mathcal{J}$ = 3												
True x^*	0.002	0.058	0.934	0.996	0.002	0.058	0.944	1.000	0.000	0.041	0.944	1.000
LW: Weighted	−0.022	0.062	0.872	0.996					0.000	0.041	0.950	1.000
YJL: Weights ∈ [0, 1]	−0.036	0.062	0.866	0.994	0.012	0.059	0.952	0.994	−0.001	0.041	0.942	0.998
PLS Index (Std)	0.001	0.060	0.934	0.988	0.001	0.060	0.934	0.988	−0.003	0.042	0.942	1.000
YJL	−0.039	0.066	0.838	0.988	0.011	0.059	0.954	0.988	0.000	0.042	0.942	1.000
PLS Index (Non-Std)	0.001	0.059	0.938	0.988	0.001	0.059	0.938	0.988	−0.003	0.041	0.948	1.000
LW	−0.040	0.066	0.854	0.984					0.000	0.041	0.944	1.000
Equal Weight Index	−0.039	0.066	0.850	0.984	−0.007	0.061	0.934	0.984	0.000	0.041	0.944	1.000
EFA Index	0.048	0.088	0.936	0.984	−0.006	0.061	0.940	0.984	0.000	0.042	0.944	1.000
PCA Index (Non-Std)	−0.128	0.132	0.006	0.984	−0.006	0.061	0.934	0.984	−0.001	0.042	0.942	1.000
Mean z-score Index	0.049	0.089	0.874	0.982	−0.006	0.061	0.940	0.982	0.000	0.042	0.946	1.000
PCA Index (Std)	−0.078	0.089	0.526	0.982	−0.007	0.061	0.942	0.982	0.000	0.042	0.944	1.000
IV ($z_1, z_2 \rightarrow z_3$)	0.103	0.147	0.806	0.960	0.249	0.289	0.538	0.960	−0.071	0.095	0.760	0.808
IV ($z_1, z_3 \rightarrow z_2$)	0.104	0.151	0.812	0.954	0.250	0.293	0.540	0.954	−0.072	0.097	0.778	0.828
IV ($z_2, z_3 \rightarrow z_1$)	0.100	0.148	0.818	0.950	0.245	0.288	0.550	0.950	−0.069	0.094	0.772	0.816
Panel B. $\mathcal{J}$ = 5												
True x^*	0.002	0.056	0.946	0.992	0.002	0.056	0.970	1.000	−0.001	0.037	0.970	1.000
PLS Index (Std)	−0.028	0.065	0.922	0.974	−0.028	0.065	0.922	0.974	0.015	0.041	0.952	1.000
LW: Weighted	−0.049	0.075	0.784	0.972					0.018	0.042	0.946	1.000
PLS Index (Non-Std)	−0.028	0.066	0.920	0.970	−0.028	0.066	0.920	0.970	0.014	0.040	0.952	1.000
LW	−0.061	0.080	0.746	0.968					0.017	0.041	0.950	1.000
YJL: Weights ∈ [0, 1]	−0.059	0.078	0.754	0.968	−0.020	0.061	0.938	0.968	0.016	0.041	0.940	0.998
YJL	−0.060	0.079	0.746	0.968	−0.020	0.061	0.944	0.968	0.017	0.042	0.946	1.000
Mean z-score Index	0.017	0.074	0.934	0.966	−0.032	0.067	0.896	0.966	0.017	0.042	0.952	1.000
Equal Weight Index	−0.061	0.080	0.742	0.966	−0.033	0.068	0.898	0.966	0.016	0.041	0.952	1.000
PCA Index (Std)	−0.096	0.105	0.304	0.964	−0.032	0.068	0.896	0.964	0.016	0.041	0.954	1.000
EFA Index	0.016	0.074	0.920	0.964	−0.033	0.067	0.904	0.964	0.017	0.041	0.950	1.000
PCA Index (Non-Std)	−0.142	0.145	0.004	0.962	−0.034	0.068	0.886	0.962	0.016	0.041	0.950	1.000
IV ($z_1, z_2 \rightarrow z_3$)	0.066	0.121	0.902	0.926	0.196	0.242	0.702	0.930	−0.047	0.075	0.890	0.922
IV ($z_2, z_3 \rightarrow z_1$)	0.067	0.121	0.910	0.924	0.198	0.243	0.690	0.924	−0.047	0.074	0.894	0.912
IV ($z_1, z_3 \rightarrow z_2$)	0.062	0.119	0.902	0.916	0.189	0.237	0.696	0.916	−0.044	0.072	0.902	0.910
Panel C. $\mathcal{J}$ = 20												
True x^*	0.000	0.064	0.942	0.968	0.000	0.064	0.936	1.000	0.002	0.045	0.936	1.000
YJL: Weights ∈ [0, 1]	−0.085	0.098	0.586	0.918	−0.050	0.075	0.886	0.918	0.040	0.057	0.848	1.000
LW: Weighted	−0.078	0.095	0.614	0.910					0.041	0.058	0.842	1.000
PLS Index (Std)	−0.058	0.083	0.814	0.906	−0.058	0.083	0.814	0.906	0.039	0.057	0.854	1.000
PLS Index (Non-Std)	−0.060	0.084	0.810	0.900	−0.060	0.084	0.810	0.900	0.040	0.057	0.848	1.000
Equal Weight Index	−0.087	0.101	0.580	0.898	−0.062	0.085	0.802	0.898	0.041	0.058	0.846	1.000
Mean z-score Index	−0.020	0.075	0.936	0.898	−0.062	0.085	0.808	0.898	0.041	0.058	0.844	1.000
PCA Index (Std)	−0.118	0.125	0.180	0.896	−0.063	0.086	0.798	0.896	0.042	0.058	0.844	1.000
EFA Index	−0.021	0.076	0.810	0.894	−0.063	0.087	0.794	0.894	0.042	0.059	0.826	1.000
LW	−0.087	0.101	0.578	0.892					0.042	0.058	0.840	1.000
YJL	−0.087	0.101	0.578	0.892	−0.051	0.076	0.904	0.892	0.042	0.059	0.828	1.000
PCA Index (Non-Std)	−0.157	0.160	0.000	0.888	−0.063	0.087	0.790	0.888	0.042	0.059	0.834	1.000
IV ($z_2, z_3 \rightarrow z_1$)	0.023	0.098	0.960	0.852	0.136	0.190	0.828	0.854	−0.014	0.060	0.946	0.966
IV ($z_1, z_2 \rightarrow z_3$)	0.018	0.098	0.942	0.830	0.129	0.187	0.860	0.832	−0.012	0.059	0.946	0.970
IV ($z_1, z_3 \rightarrow z_2$)	0.019	0.101	0.940	0.824	0.130	0.189	0.830	0.826	−0.012	0.060	0.942	0.968

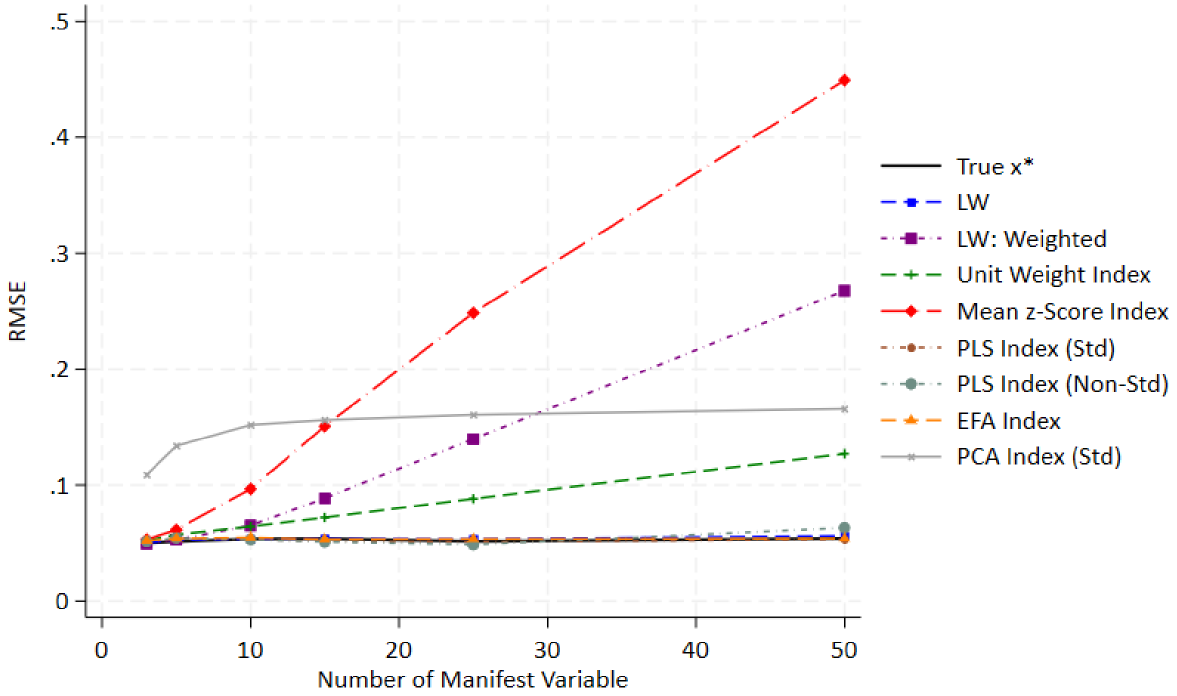
Notes.—x^* is now determined by z_j, $j = 1, ..., \mathcal{J}$, according to the formative model. Estimation based on only z_1, z_2, and z_3. Panels sorted by power of $\widehat{\beta}$.

Table J.8

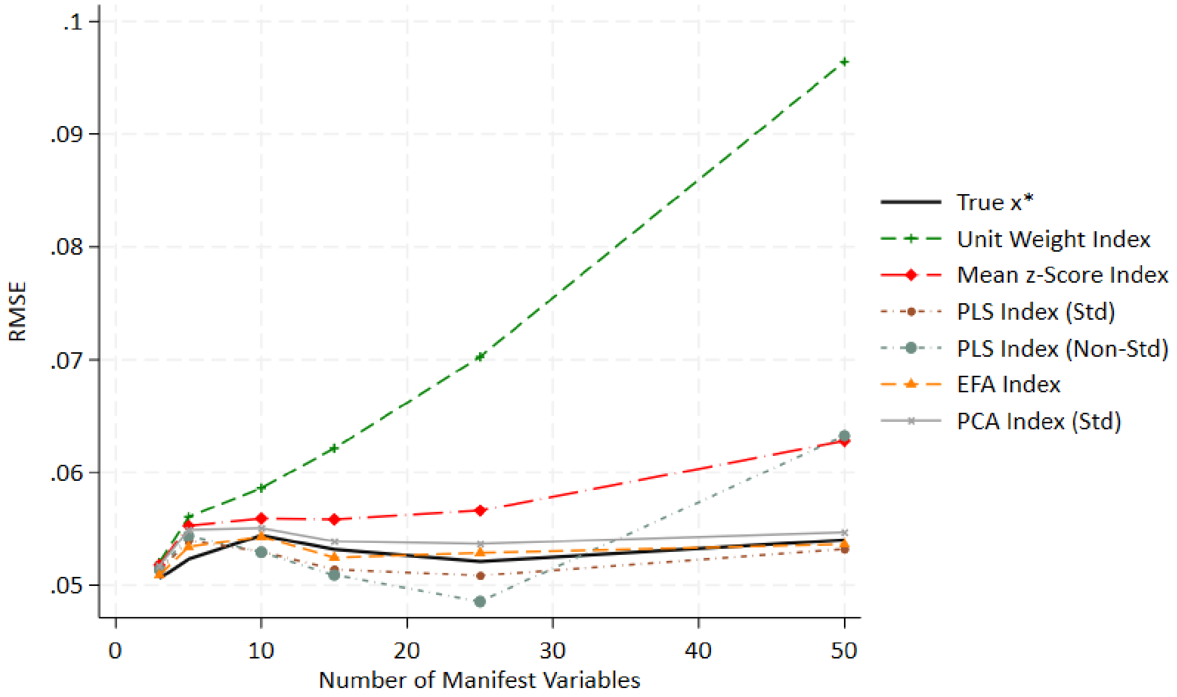
Scenario 5: Formative-Indicators Model With An Additional Exogenous Covariate ($\beta = 0.25$, $\gamma = 0$)

Method	Bias (γ)	RMSE (γ)	Coverage (γ)	False Positive (γ)	False Negative (γ)
Panel A. $\mathcal{J} = 3$					
True x^*	0.000	0.041	0.946	0.022	0.032
YJL: Weights $\in [0, 1]$	−0.001	0.040	0.950	0.020	0.030
YJL	0.000	0.041	0.942	0.028	0.030
PLS Index (Std)	−0.004	0.041	0.950	0.020	0.030
Mean z-score Index	0.000	0.041	0.946	0.024	0.030
LW	0.001	0.041	0.944	0.026	0.030
PCA Index (Non-Std)	0.001	0.041	0.944	0.028	0.028
LW: Weighted	−0.001	0.041	0.946	0.022	0.032
EFA Index	0.001	0.042	0.942	0.028	0.030
PLS Index (Non-Std)	−0.004	0.042	0.942	0.020	0.038
Equal Weight Index	−0.001	0.042	0.942	0.026	0.032
PCA Index (Std)	0.000	0.042	0.942	0.026	0.032
IV ($z_2, z_3 \rightarrow z_1$)	−0.069	0.093	0.772	0.002	0.226
IV ($z_1, z_2 \rightarrow z_3$)	−0.070	0.095	0.758	0.002	0.240
IV ($z_1, z_3 \rightarrow z_2$)	−0.071	0.096	0.784	0.000	0.216
Panel B. $\mathcal{J} = 5$					
True x^*	−0.001	0.038	0.966	0.014	0.020
PLS Index (Std)	0.013	0.040	0.958	0.036	0.006
PLS Index (Non-Std)	0.014	0.041	0.954	0.040	0.006
YJL: Weights $\in [0, 1]$	0.015	0.041	0.946	0.046	0.008
Mean z-score Index	0.017	0.041	0.948	0.046	0.006
Equal Weight Index	0.016	0.041	0.950	0.044	0.006
YJL	0.017	0.041	0.944	0.048	0.008
EFA Index	0.016	0.041	0.952	0.042	0.006
PCA Index (Non-Std)	0.017	0.042	0.950	0.044	0.006
LW	0.017	0.042	0.946	0.048	0.006
PCA Index (Std)	0.017	0.042	0.948	0.046	0.006
LW: Weighted	0.017	0.042	0.950	0.044	0.006
IV ($z_1, z_3 \rightarrow z_2$)	−0.046	0.073	0.896	0.000	0.104
IV ($z_2, z_3 \rightarrow z_1$)	−0.048	0.075	0.888	0.002	0.110
IV ($z_1, z_2 \rightarrow z_3$)	−0.048	0.075	0.886	0.002	0.112
Panel C. $\mathcal{J} = 20$					
True x^*	0.002	0.045	0.940	0.032	0.028
PLS Index (Non-Std)	0.039	0.057	0.856	0.144	0.000
PLS Index (Std)	0.039	0.057	0.856	0.144	0.000
YJL: Weights $\in [0, 1]$	0.040	0.057	0.856	0.142	0.002
Mean z-score Index	0.041	0.058	0.844	0.156	0.000
LW	0.041	0.058	0.844	0.156	0.000
LW: Weighted	0.041	0.058	0.842	0.158	0.000
YJL	0.042	0.058	0.838	0.160	0.002
PCA Index (Non-Std)	0.041	0.058	0.840	0.160	0.000
Equal Weight Index	0.042	0.059	0.844	0.156	0.000
PCA Index (Std)	0.042	0.059	0.838	0.162	0.000
EFA Index	0.042	0.059	0.824	0.176	0.000
IV ($z_1, z_2 \rightarrow z_3$)	−0.013	0.059	0.946	0.016	0.038
IV ($z_1, z_3 \rightarrow z_2$)	−0.011	0.060	0.942	0.020	0.038
IV ($z_2, z_3 \rightarrow z_1$)	−0.014	0.060	0.944	0.016	0.040

Notes.—See Table J.7. Panels sorted by RMSE of $\widehat{\gamma}$.



(A) β

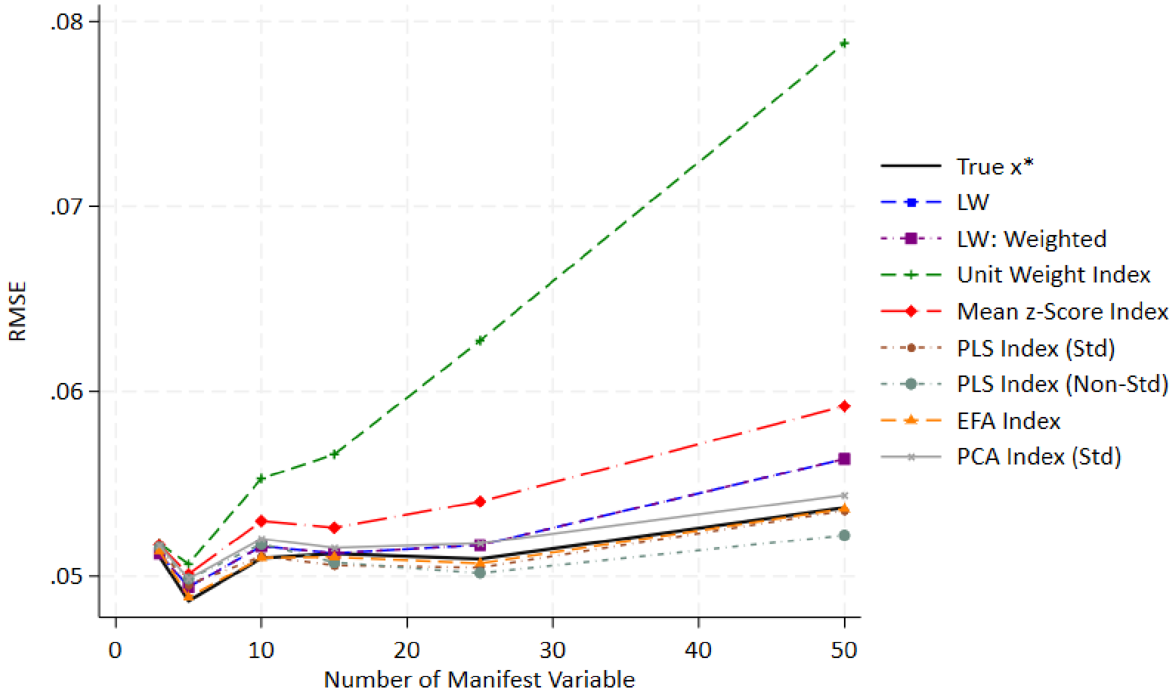


(B) η

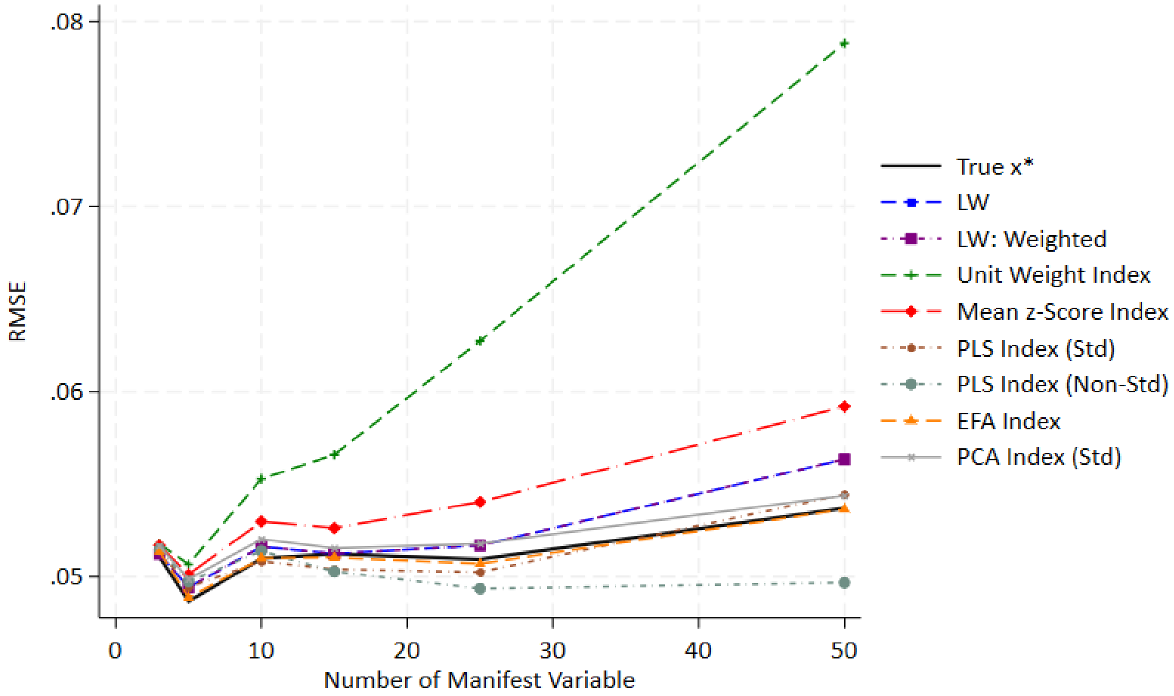
Figure J.1

Scenario 4: RMSE for β and η as the Number of Manifest Variables Vary

Notes.—See Table J.3 except the number of observed manifest variables varies. $N = 500$. The RMSE for the Mean z-score Index is sometimes suppressed when it is quite large for presentation purposes.



(A) $\gamma = 0.25$

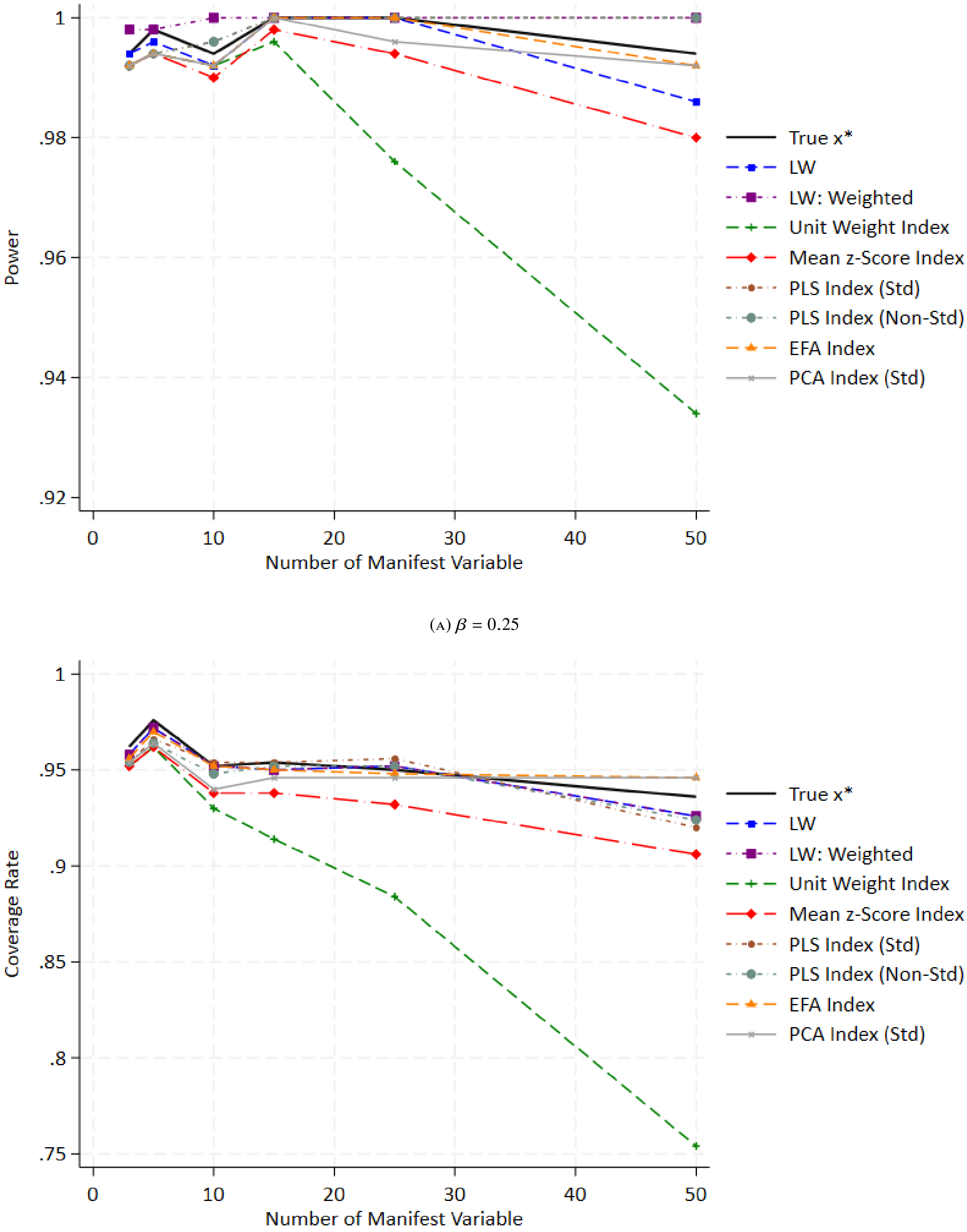


(B) $\gamma = 0$

Figure J.2

Scenario 4: RMSE for γ as the Number of Manifest Variables Vary

Notes.—See Table J.3 except the number of observed manifest variables varies. $N = 500$. The RMSE for the Mean z-score Index is sometimes suppressed when it is quite large for presentation purposes.



(A) $\beta = 0.25$

(B) $\gamma = 0$

Figure J.3
Scenario 4: Power for $\widehat{\beta}$ and Coverage Rates for $\widehat{\gamma}$ as the Number of Manifest Variables Vary ($\beta = 0.25$, $\gamma = 0$)

Notes.—See Table J.3 except the number of observed manifest variables varies. $N = 500$. True value is $\gamma = 0$ so that the coverage rate reflects the size of the test (i.e., $1 - \Pr(\text{Type I error})$).