

Rethinking Index Creation and the Usefulness of Principal Component Analysis*

Daniel L. Millimet[†]

Alfredo R. Paloyo[‡]

June 20, 2025

PRELIMINARY DRAFT

ABSTRACT

Empirical researchers often build composite indices from multiple indicators to proxy latent constructs—such as economic freedom, institutional quality, environmental stringency, or financial literacy—for use in regression analysis. Principal component analysis (PCA), although popular for constructing such indices, can lead to biased regression estimates. This paper critiques PCA's use in linear regressions, demonstrating its misalignment with consistent parameter estimation due to its focus on maximizing total variance, including noise, rather than minimizing measurement error. We show analytically and via simulations that PCA can produce greater attenuation bias than simpler alternatives. We compare PCA with methods more suited for regression, including (i) incorporating multiple indicators directly into the regression, (ii) using optimal weighting schemes from measurement-error models, and (iii) latent-factor modeling. Simulations confirm that these alternatives outperform PCA in terms of bias and mean squared error. Researchers should exercise caution with PCA-derived indices and adopt methods that explicitly address measurement error to improve inference and policy analysis.

JEL: C18; C43; C52

Keywords: principal component analysis; measurement error; reliability ratio; composite index; errors-in-variables

*The authors thank seminar participants at Charles University, SGH Warsaw School of Economics, and the University of Sydney.

[†]Southern Methodist University and IZA. Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496, United States. millimet@smu.edu.

[‡]University of Wollongong, IZA, RWI, and Crawford School of Public Policy. School of Business, University of Wollongong, Keiraville, NSW 2522, Australia. alfredo@paloyo.net.

*"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

Composite indices are ubiquitous in applied economics and other disciplines. Researchers frequently aggregate a set of observed variables—referred to as indicator or manifest variables—into a scalar index when the underlying concept of interest is multidimensional or not directly observable. Often, the concept being represented is incapable of being measured accurately as it is inherently latent. Dijkstra (2010, p. 25) states that such *theoretical* concepts are “fundamental to the scientific enterprise in almost any field”. Examples include objects such as economic freedom, institutional quality, political stability, development, regulatory stringency, human capital, personality, social capital, and financial literacy (see, e.g., Ravallion 2012). After combining various indicators, the index serves as a proxy for the unobserved latent construct, which we denote by x^* , that one wishes to include as a covariate in a regression model.¹

Focusing on linear models, the typical application relying on a composite index involves a regression of the form

$$y = \alpha + \beta x^* + \mathbf{w}'\boldsymbol{\gamma} + \varepsilon, \quad (1)$$

where y is an outcome of interest, x^* is the latent covariate (the “true” index value, e.g., the true level of institutional quality), \mathbf{w} is a vector of additional observed control variables, and ε is the error term. The researcher’s objective may be to estimate β , or β may simply be a nuisance parameter with the goal being to estimate $\boldsymbol{\gamma}$ (or one of its elements). If interest is in β , it is important to recognize that the latent construct, x^* , has no intrinsic scale. As a result, the interpretation of β is unclear. To circumvent this issue, it is perhaps most natural to conceptualize x^* as being standardized so that β is the marginal effect of a one-standard-deviation increase in the latent variable.

Although x^* is unobserved, the researcher may observe a vector of J indicators, z_1, z_2, \dots, z_J —often called “manifest” variables in the context of structural-equation modeling—where each element conveys some useful information about x^* . The challenge—that is, the *index-creation problem*—is to aggregate $\mathbf{z} = [z_1 \ z_2 \ \dots \ z_J]$ into a scalar, $x := g(\mathbf{z})$, by choosing $g(\cdot)$ such that x is a useful proxy for x^* in Equation (1). There are many ways to perform such aggregation, and the most prevalent method varies across disciplines. Our objective in this paper is to examine the consequences of researcher decisions regarding index creation.

Perhaps the most popular approach, particularly in economics, is principal component analysis (PCA), and thus we focus much of our attention on it here. Researchers commonly employ PCA for dimensionality reduction and to create indices from multiple indicators of underlying, unmeasurable

¹ Composite indices are used for non-regression purposes such as rankings of local “development”, “health”, or “welfare”. The Human Development Index (HDI) is a well-known example. Ravallion (2012) provides a thorough discussion. Here, we focus on the use of composite indices in regression analysis.

latent constructs. More explicitly, PCA generates a weighted linear combination of variables given by

$$x^{\text{PCA}} := \sum_{j=1}^J v_j z_j, \quad (2)$$

where the weights v_j are chosen to *maximize the variance* of x^{PCA} .² Since PCA finds linear combinations that “successively maximize variance” without reference to any outcome variable, it offers a seemingly objective way to aggregate many correlated indicators into a scalar index (Jolliffe and Cadima 2016, p. 1). PCA’s dimensionality reduction and weighting based on observed variation have made it the default tool for index construction across many fields.³

Yet, from an econometric standpoint, PCA is generally suboptimal for regression purposes. The core issue is that maximizing the total variance of the index, x , is not the same as maximizing the signal-to-noise ratio with respect to the latent variable of interest. This property compromises the reliability of policy conclusions derived from PCA-based indices. In this paper, we illustrate the extent of this issue analytically and empirically, emphasizing the econometric risks of ignoring the inevitable measurement error associated with PCA-based index construction.

Consider the case of classical measurement error in an observed index, x , used in place of the “true” but unobserved construct, x^* . In this case, it is well known that the reliability ratio of the index—that is, the fraction of the index’s (partial) variance that is due to the true signal x^* rather than noise—determines the extent of attenuation bias in regression estimates obtained using ordinary least squares (OLS).⁴ As such, the “optimal” index for regression purposes is the one that maximizes reliability (or, equivalently, minimizes the variance of measurement error in the index x). PCA, however, ignores the distinction between common variance (signal) and idiosyncratic variance (noise) in \mathbf{z} . By construction, PCA will *overweight* noisy components in \mathbf{z} if those components exhibit high total variability. This can lead to an index with a lower reliability ratio—and hence more severe attenuation bias—than alternative weighting schemes or even a simple average of the manifest variables. In short, OLS relying on PCA does not generally yield the estimate of β in Equation (1) with the smallest bias.

We are not the first to recognize this shortcoming of PCA. Lubotsky and Wittenberg (2006, p. 551) state:

Principal-component analysis achieves a unique decomposition, but does so by the expedient of identifying the common factor with the linear combination of indicators that maximizes the explained variance. It is not clear why this concept should correspond to the structural relationships. . .

² Formally, PCA can be used to generate several aggregates. The so-called first principal component chooses the weights, v_j , $j = 1, 2, \dots, J$ to maximize the variance of x^{PCA} . The second (and higher) principal component chooses the weights, $v_j^{(2)}$, $j = 1, 2, \dots, J$ to maximize the variance of $x^{\text{PCA}(2)}$ subject to the restriction that $x^{\text{PCA}(2)}$ (or higher principal components) is orthogonal to x^{PCA} (and any other previous principal components). As nearly all researchers rely solely on the first principal component (PC1) as the index, we simplify the exposition by implicitly referring to PC1 throughout the paper.

³ A search on Google Scholar for the terms “‘principal component analysis’ AND ‘index’ AND ‘regression’” yields 648,000 hits; 19,700 since 2024. Accessed on 01 April 2025.

⁴ As discussed in Section 2, the degree of attenuation depends on the signal-to-noise ratio in the mismeasured covariate *net of all other covariates in the model*.

Going back even further in time, economists recognized the inherent arbitrariness in the construction of most composite indices. Mundlak (1961, p. 44) denounces the use of an index of farmer managerial skill when estimating production functions, writing:

It has been felt for a long time that the estimates of the parameters of production functions are subject to bias as a result of excluding the variable which represents management. The reason for omitting management from cross-section analysis is obviously the lack of units for its direct measurement. An attempt to substitute some index of management does not solve the conceptual difficulty. It can be regarded as an ad hoc procedure as long as no criterion for evaluating its performance is available.

Similarly, Samuelson (1983, p. 144)⁵ urges caution in the context of aggregating commodities, noting:

There is nothing intrinsically reprehensible in working with such aggregate concepts. On the contrary, abstraction from complexity is a necessary thought process. . . . But it is important to realize the limitations of these aggregates and to analyze the nature of their construction.

More recently, Ravallion (2012, p. 2) writes:

[T]he analyst identifies a set of indicators that are assumed to reflect various dimensions of some unobserved (theoretical) concept. An aggregate index is then constructed. Neither the menu of the primary series nor the aggregation function is predetermined from theory and practice, but are “moving parts” of the index—key decision variables that the analyst is free to choose, largely unconstrained by economic or other theories intended to inform measurement practice.

The author refers to indices that lack a theoretical or even practical foundation as “mashup indices”.

Despite these warnings, the shortcomings of composite indices in general and PCA in particular have gone largely unrecognized by applied researchers—at least from our perspective. However, the choices being made by researchers have substantive implications. Estimates of β (the effect of the latent variable) using a PCA index will be biased toward zero and inconsistent assuming classical measurement error (Griliches 1977; Hyslop and Imbens 2001). The severity of the bias can be significant, particularly when the manifest variables vary in quality. Our simulations demonstrate that PCA-based indices can underestimate true parameter values by XX% to YY%, potentially leading to incorrect policy conclusions.

We emphasize that many applied studies using PCA indices may be understating or mismeasuring the true impact of the latent concept on outcomes. For example, if institutional quality truly has a large effect on economic growth, but is proxied by a PCA index built from imperfect governance indicators, the estimated coefficient on the PCA index could be greatly attenuated. Policy decisions based on such results might underweight the importance of institutional improvements. Thus, assessing and improving index construction is not a mere technical exercise, but vital for credible inference.

Even if the researcher has no interest in β itself, but rather includes the index as a necessary control variable in an effort to obtain unbiased estimates of γ (trying to avoid omitted-variable bias),

⁵ Accessed at <https://archive.org/details/in.ernet.dli.2015.150369/page/n147/mode/2up> on 01 April 2025.

measurement error in the index matters if the partial correlation between x and any element of \mathbf{w} is nonzero. In this situation, the bias arising from the ignored endogeneity of x (due to measurement error) will bias the estimate of γ as well (see, e.g., Hanushek and Jackson 1977; Griliches 1986; Bollen 1989). Bollinger (2003, p. 578) discusses the use of proxies in linear regression more generally, stating that a “maintained, but incorrect, assumption is that the coefficients on other variables are identified correctly when this proxy variable is used.” The author concludes (p. 583) by offering a “general overall warning to researchers: when proxy variables are used, coefficients on other variables may also be biased—in some cases severely.”⁶

Our paper is most similar to Stoetzer, Zhou, and Steenbergen (2025). The authors make the same point as we do here in the context of situations where the latent construct is the *outcome of interest* when estimating the average treatment effect of an intervention that satisfies unconditional or conditional unconfoundedness. Not only does this situation raise issues with the outcome of interest being latent, but also the interpretation of the magnitude of the average treatment effect since the latent construct has no intrinsic scale. The authors show that two-step approaches that estimate the latent construct using PCA or alternatives in the first-stage and the relative average treatment effect (expressed in standard deviations of the latent construct) in the second stage will generally be biased. The authors propose a hierarchical item response theory model that avoids direct estimation of the latent construct by simultaneously estimating parametric equations for the latent construct and the observed indicators using an EM algorithm. Although extremely useful, the focus on the estimation of the so-called latent average treatment effect is limiting. Our analysis broadens the discussion to situations where the latent construct is a covariate in a (linear) regression model and its coefficient may or may not be the object of interest.

Our paper is also quite similar to Filmer and Scott (2012). The authors assess eight different methods for computing a household asset index and compare the relative rankings of households across the different indices as well as the characteristics of households across the quintiles of each index. We undertake a similar comparison of multiple approaches to index creation, but in the context of linear regression.

(Describe next sections.)

In this paper, we contribute a rigorous examination of PCA’s limitations in index construction for regressions, and we explore better alternatives. We extend previous critiques by providing analytical derivations, simulations across multiple realistic scenarios, and clear methodological guidelines to practically address and correct measurement-error issues associated with index construction.

In Section 6, we provide simulation evidence to illustrate the comparative performance of PCA and the proposed alternatives. We design Monte Carlo experiments where the true latent variable x^* and multiple indicators are generated with distinct error variances. We evaluate each method’s ability to recover the true β and γ using bias, coverage probability of the nominal 95% confidence intervals for β , and root mean square error as criteria. The simulations show that PCA often performs worse than even “naïvely” including all indicators in the regression. In contrast, methods that explicitly account for measurement error (such as instrumental variables (IV) estimation) produce estimates of β that are

⁶ Recently, Zhang and Lee (2025) provide a somewhat general characterization of when inclusion of a mismeasured covariate is preferred over its exclusion in the case where γ corresponds to the average treatment effect of a binary intervention.

much closer to the true value. We also propose a simple method—implemented via generalized method of moments (GMM)—to estimate the optimal index weights.

Finally, Section 9 concludes with a discussion of implications for applied work. We argue that researchers should exercise caution when using PCA-derived indices as regressors: one should assess the index’s reliability and consider alternative construction methods. Where feasible, including multiple indicator variables directly in regression or employing measurement-error correction techniques can substantially reduce bias. We also offer some practical guidance on index construction, recommending transparency (e.g., reporting the variance explained by the index and performing robustness checks with alternative weighting schemes). The broader message is that not all indices are created equal: methods that align with econometric objectives of consistent estimation should be preferred over mechanically convenient but potentially distortionary tools like PCA.

2 Index Creation

2.1 Formal Setup for Index Creation

We consider the regression framework of Equation (1), where x^* is a latent scalar covariate. Researchers observe a vector of J manifest or indicator variables, z_1, z_2, \dots, z_J , that contain information about x^* . Although not typically discussed by researchers, there are two distinct ways to conceptualize the relationship between the indicators and the latent variable (e.g., Stoetzer, Zhou, and Steenbergen 2025). These are illustrated in Figure 1.

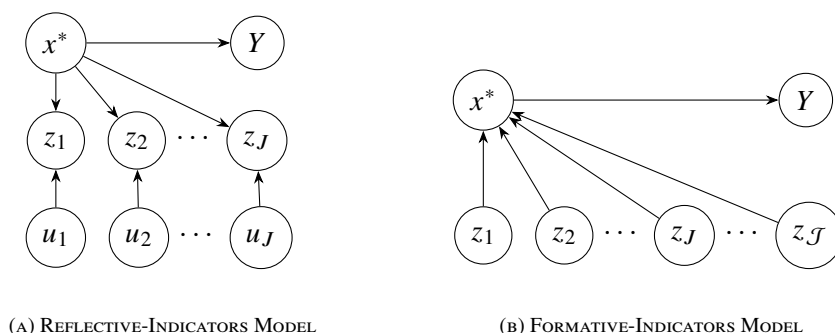


FIGURE 1
RELATIONSHIP BETWEEN INDICATORS AND LATENT VARIABLE

NOTES.— Blah blah blah.

Panel A displays the *reflective-indicators model*. In this case, the indicators “reflect” the latent variable, x^* , as well as other unmodeled attributes, denoted by u . Panel B reflects an alternative conceptualization referred to as the *formative-indicators model*. Here, the indicators are determinants that help “form” x^* . In the reflective model, measurement error arises due to the presence of u . In the formative model, measurement error arises due to the presence of additional indicators, denoted as z_j , $j = J + 1, \dots, \mathcal{J}$, not observed by the researcher. An example of the reflective model is in Millimet, McDonough, and Fomby (2018), where the answers to several financial questions are used to create an

index of an individual's financial literacy. In this case, "true" financial literacy determines the correctness of the answers along with other unobserved factors such as how seriously the respondent took the survey. An example of the formative model is in Maccini and Yang (2009) and Filmer and Scott (2012) where the presence of many different types of assets are used to construct a household asset index. Here, the z 's represent various assets (e.g., auto, home, television, investments, etc.) and x^* is latent aggregate of these assets. To proceed, we focus on the reflective-indicators model. **We will return to the formative-indicators model later.** In our view, the reflective model is the way researchers typically conceive of the relationship between the indicators, z , and the latent variable, x^* . It gives rise to the traditional approach of expressing each indicator as

$$z_j = x^* + u_j, \quad j = 1, 2, \dots, J, \quad (3)$$

where u_j is classical measurement error in indicator j . Specifically, we assume $E[u_j] = 0$, $\text{Cov}(u_j, x^*) = 0$, $\text{Cov}(u_i, u_j) = 0$ for all $i \neq j$, $\text{Cov}(u_j, \varepsilon) = 0$, and $E[\tilde{\mathbf{w}}\varepsilon] = 0$, where $\tilde{\mathbf{w}} = [x^* \ \mathbf{w}]$. In other words, each z_j is an unbiased reflections of x^* , the indicators are independent conditional on x^* , and estimation of Equation (1) by OLS will produce unbiased and consistent estimates if x^* is observed. **We consider deviations from the classical measurement-error setup later.**

An index $x := g(z_1, z_2, \dots, z_J)$ is any function that aggregates the indicators into a scalar intended to approximate x^* . We abstract from the choice of the indicators themselves and take \mathbf{z} as given. **We relax this later.** Linear indices are most common in practice.⁷ These are of the form:

$$x = \sum_{j=1}^J \lambda_j z_j, \quad (4)$$

with weights λ_j chosen by the researcher. The index or proxy error is $\mu := x - x^*$. Different choices of the aggregation function $g(\cdot)$ will generally lead to different values of μ and hence different levels of similarity between x and x^* .

One convenient measure of index "quality" is the reliability ratio, defined as

$$\rho := \frac{\text{Var}(x^*)}{\text{Var}(x)} = \frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(\mu)} = 1 - \frac{\text{Var}(\mu)}{\text{Var}(x)}. \quad (5)$$

The reliability ratio ρ is the fraction of the index's variance coming from the true latent signal rather than noise. Equivalently, $\text{Var}(\mu) = (1 - \rho) \text{Var}(x)$, so a higher reliability ratio ρ implies lower error variance. An index with $\rho = 1$ perfectly measures x^* ; an index with $\rho = 0$ is pure noise. The relevance of ρ is that it directly governs attenuation bias in the regression of y on x in classical settings. Substituting $x^* = x - \mu$ into Equation (1), we get

$$y = \alpha + \beta(x - \mu) + \mathbf{w}'\boldsymbol{\gamma} + \varepsilon = \alpha + \beta x + \mathbf{w}'\boldsymbol{\gamma} + (\varepsilon - \beta\mu), \quad (6)$$

where $(\varepsilon - \beta\mu)$ is the composite error. Here, x is observed, but the regression error term now includes $-\beta\mu$, which is generally correlated with x (since μ is part of x), thus violating the OLS orthogonality

⁷ There is no theoretical reason to only consider the class of linear indices. However, applied researchers always do, to our knowledge, so we leave consideration of nonlinear indices for future work.

condition.

In a univariate regression model, the probability limit (plim) of the OLS slope estimate using a mismeasured regressor is $\beta\rho$. The coefficient is attenuated by a factor equal to the reliability ratio. In a multiple regression where only the index suffers from classical measurement error, the plim of the OLS coefficient estimate for the incorrectly measured covariate is

$$\text{plim} \left[\widehat{\beta} \right] = \beta \left[1 - \frac{\text{Var}(\mu)}{\text{Var}(x) (1 - R_{x,w}^2)} \right], \quad (7)$$

where $R_{x,w}^2$ is the R^2 from the regression of x on \mathbf{w} (Griliches 1977). Equation (7) reduces to $\beta\rho$ when x and \mathbf{w} are orthogonal. If x and \mathbf{w} are not orthogonal, which occurs if \mathbf{w} is correlated with x^* and/or μ , then not only is the attenuation bias exacerbated, but the OLS estimate of γ is also biased (Hanushek and Jackson 1977; Griliches 1986; Bollinger 2003; Bollinger and Minier 2015). The bias in γ may be in any direction.

In sum, when using an index x in lieu of the unobserved x^* , OLS consistently estimates β only if $\rho = 1$ (i.e., no measurement error) and γ only if $\rho = 1$ and/or x and \mathbf{w} are orthogonal. Moreover, the coefficient on the proxy is biased toward zero in the classical setting, with severity increasing as ρ declines and $R_{x,w}^2$ increases. Thus, the “optimal” index for inclusion in regression analysis where the objective is consistent estimation of the slope parameters is the one that maximizes ρ . Maximizing ρ (or, equivalently, minimizing $\text{Var}(\mu)$) should then guide a researcher’s choice of the aggregation function, $g(\cdot)$, in general and the linear weights, λ_j , $j = 1, \dots, J$, in Equation (4) in particular. Intuitively, x should track x^* as closely as possible.

In the simple classical setting considered thus far, where the measurement error is independent across indicators (as in Panel A in Figure 1), the instrumental variables (IV) estimator—where z_{-j} is used as an instrument for z_j —is consistent for β and γ . While we discuss the IV estimator below, we consider the ramifications of the choice of different indices since the IV estimator is inconsistent when the measurement error is correlated across the indicators. Correlated measurement errors are likely to be the norm in practice; the exposition above is limited to the classical setting for simplicity only.

2.2 Optimal Linear Index Creation

Consider the index $x = \sum_j \lambda_j z_j$ and its reliability ratio $\rho = \text{Var}(x^*) / \text{Var}(x)$ from Equation (4). Maintaining the assumptions from the previous section, the variance of a generic linear index is

$$\text{Var}(x) = \text{Var} \left(\sum_j \lambda_j z_j \right) = \text{Var} \left(\sum_j \lambda_j x^* + \sum_j \lambda_j u_j \right) = \left(\sum_j \lambda_j \right)^2 \sigma_{x^*}^2 + \sum_j \lambda_j^2 \sigma_{u_j}^2 \quad (8)$$

where $\text{Var}(x^*) = \sigma_{x^*}^2$ and $\text{Var}(u_j) = \sigma_{u_j}^2$. For simplicity, we impose the normalization $\sum_j \lambda_j = 1$ so that the index is an unbiased estimator of x^* . This constraint fixes the scale of λ . It also ensures that x equals x^* in the absence of measurement error in the indicators. Under this normalization,

$\text{Var}(x) = \sigma_{x^*}^2 + \sum_j \lambda_j^2 \sigma_{u_j}^2$. The reliability ratio is given by

$$\rho(\lambda) := \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sum_j \lambda_j^2 \sigma_{u_j}^2}. \quad (9)$$

The optimal index should maximize ρ , which is equivalent to minimizing the total error variance $\sum_j \lambda_j^2 \sigma_{u_j}^2$ subject to $\sum_j \lambda_j = 1$. This is a straightforward constrained optimization problem. One can show the solution is

$$\lambda_j^* \propto \frac{1}{\sigma_{u_j}^2}, \quad j = 1, \dots, J, \quad (10)$$

that is, weights are proportional to the inverse of an indicator's error variance (assuming $\sigma_{u_j}^2 > 0$ for all j). Intuitively, the most reliable indicator receives the greatest weight, and the noisiest indicator, the least. The resulting optimal index $x^* = \sum_j \lambda_j^* z_j$ has the highest reliability among all linear combinations of \mathbf{z} under the given normalization. The reliability ratio of x^* is obtained by substituting Equation (10) into Equation (9). In general, as more independent indicators are added, the reliability ratio rises, approaching one as $J \rightarrow \infty$ (Lubotsky and Wittenberg 2006). With a finite number of indicators, the reliability ratio remains strictly below one, and thus even the optimal linear index yields some attenuation bias—but crucially, it minimizes that bias.

In the absence of validation data, which is always the case in situations where x^* denotes a theoretical construct that is inherently unobservable (such as freedom, development, or health), the optimal weights in Equation (10) are unknown. Nonetheless, this discussion provides a framework in which to assess PCA-constructed indices and alternatives, as we discuss next.

3 The Case Against PCA in Applied Regression Analysis

The fundamental econometric drawback of PCA-constructed indices is their failure to minimize measurement error. In the language of classical errors-in-variables, PCA does not maximize the reliability ratio ρ of the proxy. As a result, using a PCA index in place of x^* in Equation (1) will render OLS biased, perhaps severely so, as an estimator of β —an issue that remains even as sample size grows since the bias is due to systematic error in the regressor (Dong and Millimet 2024).

3.1 Principal Component Analysis for Index Creation

PCA finds linear combinations of the input variables that successively maximize variance. In the context of index creation, PCA will produce the index (first principal component or PC1) $x^{\text{PCA}} = \sum_{j=1}^J v_j z_j$ that has the largest variance among all unit-length weighting vectors $\mathbf{v} = [v_1 \ v_2 \ \dots \ v_J]$. This is obtained by solving

$$\mathbf{v}^{\text{PCA}} = \arg \max_{\|\mathbf{v}\|=1} \text{Var}(\mathbf{v}'\mathbf{z}), \quad (11)$$

which leads to \mathbf{v}^{PCA} being the eigenvector associated with the largest eigenvalue of $\text{Var}(\mathbf{z})$, the covariance matrix of the z_j 's. Imposing the unit norm $\|\mathbf{v}\| = 1$ fixes the scale since, otherwise, one could increase variance arbitrarily by scaling up \mathbf{v} . The resulting index $x^{\text{PCA}} = \mathbf{v}^{\text{PCA}'} \mathbf{z}$ has variance equal to the largest

eigenvalue of $\text{Var}(\mathbf{z})$. In practice, researchers often standardize the variables z_j before applying PCA so that $\text{Var}(z_j) = 1$ for all j , especially if the indicators are measured in different units. PCA on standardized data finds the linear combination with maximum correlation or normalized variance.

The PCA index is attractive for several reasons. First, it provides dimensionality reduction: instead of J separate variables, one uses a single index that, by construction, captures as much of their joint variation as possible. Second, PCA uses data-driven weights v_j , often interpreted as “letting the data speak” about which indicators are most important. Variables with more variability receive larger weights in the first component, which some have taken as a feature. Third, the resulting principal components are uncorrelated with each other, which, for the sake of tradition and those inexplicably still focused on it, addresses multicollinearity concerns if one were to use multiple indices or subsequent components. These properties explain PCA’s popularity: it yields a single summary index, avoids subjective weighting, and handles collinearity by construction.

However, it is crucial to note what PCA does not consider: any information about the relationship of \mathbf{z} with the outcome y or the latent factor x^* .⁸ The criterion depends solely on the variance of \mathbf{z} . In effect, PCA treats all variation in the indicators as valuable signal regardless of whether it stems from the common latent factor or idiosyncratic noise. As a result, if one indicator z_k has a very large variance due to measurement error, PCA will tend to assign z_k a large weight—because including more of z_k increases the total variance of the index—even though z_k may be a poor measure of x^* . In contrast, an indicator with lower variance might receive a smaller PCA weight or even be omitted from the first component. This behavior is antithetical to the goal of maximizing reliability.⁹

To fix ideas, suppose two indicators z_1 and z_2 measure the same latent x^* with different noise levels. For example, $z_1 = x^* + u_1$ with $\text{Var}(u_1) = \sigma_{u_1}^2$, and $z_2 = x^* + u_2$ with $\text{Var}(u_2) = \sigma_{u_2}^2$. If $\sigma_{u_2}^2 \gg \sigma_{u_1}^2$, then z_2 exhibits greater total variance than z_1 because z_2 ’s variance $\text{Var}(z_2) = \text{Var}(x^*) + \sigma_{u_2}^2$ is much larger. PCA will favor z_2 . In fact, if the difference is extreme, the first principal component (PC1) will be almost aligned with z_2 . Consequently, the PCA index $x^{\text{PCA}} \approx z_2$ is a noisy proxy, inheriting z_2 ’s low reliability. The coefficient estimate on x^{PCA} in a regression will be severely attenuated toward zero in a univariate regression, being nearly as bad as using z_2 alone. In contrast, an index that gave more weight to z_1 —the higher-quality indicator—would achieve a higher ρ and yield a less biased estimate of β . This simple example underscores the point that maximizing variance in the index x is not equivalent to maximizing information about the latent factor x^* in a regression sense.

⁸ In the language of machine learning, PCA is an unsupervised technique: the weights v_j are chosen without reference to β or the regression model. Unsupervised machine learning algorithms analyze “unlabeled” data or do not take the labels into account. They simply try to uncover patterns or data groupings. Dimensionality-reducing techniques, like principal component analysis or singular value decomposition, are examples of unsupervised learning. In the context of regression analysis, supervised learning models would use the independent and dependent variables to discern their relationship for prediction or causal inference. As an unsupervised technique, PCA is used as a pre-processing step to reduce the dimensionality of the data—that is, to reduce the number of regressors—without regard to how the independent variables relate to the outcome of interest.

⁹ We will see in Section 3.2 that PCA’s variance-maximizing index can have a substantially lower reliability ratio ρ than an index that explicitly accounts for measurement error.

3.2 PCA Does Not Maximize the Reliability of the Proxy

PCA chooses \mathbf{v} to maximize $\text{Var}(x^{\text{PCA}}) = \left(\sum_j v_j\right)^2 \sigma_{x^*}^2 + \sum_j v_j^2 \sigma_{u_j}^2$ subject to the unit-norm constraint $\sum_j v_j^2 = 1$ rather than the unit-sum constraint $\sum_j v_j = 1$ used for the optimal index. The two objectives—maximizing variance vs. maximizing reliability—coincide only in knife-edge cases (e.g., all $\sigma_{u_j}^2$ equal or all indicators perfectly correlated), so \mathbf{v}^{PCA} generally differs from the optimal weights in Equation (10). The weights in \mathbf{v}^{PCA} are influenced by the relative magnitudes of $\sigma_{u_j}^2$ plus the presence of the common variance $\sigma_{x^*}^2$. If a manifest indicator has a very large idiosyncratic variance $\sigma_{u_j}^2$, it increases the diagonal entry of $\text{Var}(\mathbf{z})$ and can dominate the first principal component even if $\sigma_{u_j}^2$ is pure noise. The PCA index thus overweights high-variance indicators regardless of whether the variance comes from signal or noise.

Another viewpoint is to compare the reliability ratio ρ achieved by PCA versus the optimum. Denote $\rho^{\text{PCA}} = \text{Var}(x^*) / \text{Var}(x^{\text{PCA}})$ and $\rho^* = \text{Var}(x^*) / \text{Var}(x^*)$.¹⁰ Generally, $\rho^{\text{PCA}} < \rho^*$ unless PCA, by coincidence, picks the same weights as Equation (10). We can see this in a simple case of two manifest indicators, z_1 and z_2 . Let $\sigma_{x^*}^2 = 1$ without loss of generality, and let $\sigma_{u_1}^2 < \sigma_{u_2}^2$ so indicator 1 is more reliable. The optimal weights, with $\lambda_1 + \lambda_2 = 1$, are $\lambda_1^* = \sigma_{u_2}^2 / (\sigma_{u_1}^2 + \sigma_{u_2}^2)$ and $\lambda_2^* = \sigma_{u_1}^2 / (\sigma_{u_1}^2 + \sigma_{u_2}^2)$ (Lubotsky and Wittenberg 2006). Intuitively, if z_2 has a larger error variance, one weights it less. Plugging in, we obtain $\rho^* = 1 / (1 + \lambda_1^{*2} \sigma_{u_1}^2 + \lambda_2^{*2} \sigma_{u_2}^2)$. One can then verify that ρ^* is higher (closer to 1) than the reliability of using either indicator alone and is, in fact, the maximum attainable. In contrast, PCA in this 2-variable case will set \mathbf{v}^{PCA} as the eigenvector of the 2×2 variance–covariance matrix of (z_1, z_2) .

For a numerical example, consider $\sigma_{u_1}^2 = 0.5$ and $\sigma_{u_2}^2 = 4$, i.e., indicator 1 is high-quality and indicator 2 is very noisy. Then the optimal weights, normalized to sum to 1, are $\lambda_1^* \approx 0.89$ and $\lambda_2^* \approx 0.11$. The PCA weights, after first standardizing the manifest indicators (i.e., z -scores), are $v_1^{\text{PCA}} \approx 0.71$ and $v_2^{\text{PCA}} \approx 0.71$, which corresponds to the raw (non-standardized) indicators weighted with 0.26 and 0.97, respectively.¹¹ PCA puts more weight on the noisier indicator. The reliability of the PCA index can be computed in this example: here, $\rho^{\text{PCA}} \approx 0.59$ when the indicators are standardized (the reliability ratio of the PCA index based on non-standardized indicators degrades to 0.28), whereas $\rho^* \approx 0.69$. Consequently, the asymptotic attenuation bias—see Equation (7)—using the PCA index would be about 41% compared to 31% if one used the optimally weighted index. Even using the better indicator z_1 alone would have given $\rho_1 = 1 / (1 + 0.5) = 0.67$, which is very close to the optimum in this case.

In general, PCA-based weights can lead to severe measurement-error bias. The fundamental reason is that PCA’s criterion—maximize $\text{Var}(x)$ —does not consider the index or proxy error $\text{Var}(\mu)$ at all. In fact, PCA will increase $\text{Var}(x)$ by increasing $\text{Var}(\mu)$ if doing so boosts total variance. In our two-indicator example, z_2 has so much variation from noise that PCA chooses to exploit it to enlarge the index variance at the cost of injecting more error. From the perspective of estimating β and the other regression parameters, this is undesirable.

¹⁰ The generic expression for the reliability ratio is $\rho = \left[(\sum_j w_j)^2 \sigma_{x^*}^2 \right] / \left[(\sum_j w_j)^2 \sigma_{x^*}^2 + \sum_j w_j^2 \sigma_{u_j}^2 \right]$, with $w = v$ for PCA and $w = \lambda$ for the optimal index. See also Equation (9).

¹¹ The weights are equal if the indicators are standardized first because variances are equalized after standardization. Whether standardized or not, however, the weights satisfy the unit-norm constraint.

3.3 Bias and Inconsistency of PCA-Based Estimation

We now quantify the bias in OLS estimation when using a PCA index. Consider again Equation (1), but we substitute x^{PCA} for the unobserved x^* . The estimating equation becomes

$$y = \alpha + \beta x^{\text{PCA}} + \mathbf{w}'\boldsymbol{\gamma} + \tilde{\varepsilon}, \quad (12)$$

where $\tilde{\varepsilon} = \varepsilon + \beta(x^* - x^{\text{PCA}}) = \varepsilon - \beta\mu^{\text{PCA}}$. Since x^{PCA} is correlated with μ^{PCA} , the regressor becomes endogenous. In the classical setting, one can apply Equation (7) to characterize the probability limit of the OLS estimator $\hat{\beta}^{\text{PCA}}$:

$$\text{plim} \left[\hat{\beta}^{\text{PCA}} \right] = \beta \left[1 - \frac{\text{Var}(\mu^{\text{PCA}})}{\text{Var}(x^{\text{PCA}}) (1 - R_{x^{\text{PCA}}, \mathbf{w}}^2)} \right] = \beta \left[1 - \frac{1 + (1 - 2 \sum_j v_j) \rho^{\text{PCA}}}{1 - R_{x^{\text{PCA}}, \mathbf{w}}^2} \right]. \quad (13)$$

This equation reveals that because PCA employs the unit-norm constraint $\sum_j v_j^2 = 1$ (which does not generally imply $\sum_j v_j = 1$), $\hat{\beta}^{\text{PCA}}$ does not necessarily suffer from attenuation bias. For instance, if $\sum_j v_j > 1$, and PCA generates a sufficiently high reliability ratio ρ^{PCA} , $\hat{\beta}^{\text{PCA}}$ can be subject to expansion bias (biased upward in absolute value). In this scenario, the PCA weighting effectively “overcompensates” for the measurement error, leading to $\text{plim} \left[\hat{\beta}^{\text{PCA}} \right] > \beta$. Conversely, as the reliability ratio ρ^{PCA} approaches zero, $\hat{\beta}^{\text{PCA}}$ will experience severe bias, potentially including sign reversal, as long as $\sum_j v_j$ is finite. Moreover, $\hat{\beta}^{\text{PCA}}$ is consistent only if the numerator in Equation (13) is zero, which means $\rho^{\text{PCA}} = 1/(2 \sum_j v_j - 1)$. This general consistency condition simplifies to $\rho^{\text{PCA}} = 1$ if and only if $\sum_j v_j = 1$, which occurs with degenerate PCA (either a single indicator only or having multiple indicators but only one receiving the nonzero weight of 1).

These biases exist even when all indicators are valid measures of the same underlying x^* and even if PCA captures most of the total variance as the total variance includes variation due to the measurement error in the indicators. In other words, for PCA, simply having $\rho^{\text{PCA}} = 1$ is not enough for consistency unless that perfect reliability is achieved under the condition that the sum of the weights is 1, which is neither a realistic nor a useful scenario. If $\sum_j v_j \neq 1$, then even if x^{PCA} perfectly captures x^* for some reason, Equation (13) shows that the probability limit is still affected by the non-classical measurement error in x^{PCA} . For while the errors attached to the manifest indicators may be classical, it does not necessarily follow that the resulting PCA index will itself exhibit classical measurement error in general. Thus, PCA does not “solve” or even minimize the problems arising from measurement error; it is simply a particular transformation of the data. Furthermore, as mentioned previously, the use of an imperfect proxy for x^* will bias the OLS estimate of $\boldsymbol{\gamma}$ if x and \mathbf{w} are not orthogonal. In many applications, this is likely to be the case.

In light of the above, we urge extreme caution to researchers when using PCA to create a proxy for a latent variable in a typical linear regression framework when the objective is causal inference. Unless one can be confident that all indicators have identical noise levels or that idiosyncratic variances are negligible (so that PCA happens to coincide with the optimal index), relying on PCA to create a proxy is not advisable. The case against PCA is essentially an application of the classical errors-in-variables

lesson: one should use methods that explicitly address measurement error rather than ignoring it. PCA ignores it by design and, as a consequence, guarantees neither consistency nor minimum bias except under special circumstances.

4 Alternative Approaches for Index Creation

4.1 Dimensionality Reduction Without PCA

Although the optimal index weights in Equation (10) are infeasible because they depend on the variances of the unobserved measurement errors, they do feature the desirable property where higher weights are attached to indicators with lower measurement-error variance. In contrast, PCA places “too much” weight on indicators with a higher measurement-error variance. Alternative indices that use feasible weights which are also closer to the optimal weights potentially offer a significant improvement over PCA in a regression context. We consider several such alternatives.

Equal Weights. A simple alternative is the equally-weighted average of the non-standardized z_j (also referred to as unit weights).¹² This is equivalent to a count index, equal to the sum of the z_j ’s, scaled by J (Filmer and Scott 2012). The index, denoted $x^{\bar{z}}$ and referred to as the equal index, sets $\lambda_j = 1/J$ for all j . The probability limit of $\widehat{\beta^{\bar{z}}}$ is $\beta\rho^{\bar{z}}$, where $\rho^{\bar{z}} \in [0, 1]$. Thus, the equally-weighted index is consistent in the absence of measurement error and otherwise suffers from the usual attenuation bias in the classical setting.¹³ With uncorrelated errors, it has a higher reliability ratio than using any single indicator in the classical setting where $\text{Var}(u_j) = \sigma_u^2$ for all j . Unit weights are also less susceptible to outliers as the weights are chosen without reference to the data (Bobko, Roth, and Buster 2007).¹⁴ In psychology, Rönkkö, McIntosh, and Antonakis (2015, p. 77) refers 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). In our two-indicator example from Section 3.2, the $\lambda_j = 0.5$ weights correspond to a reliability ratio of 0.47.

Mean z -Score. A popular alternative to PCA is the mean z -score index (Kling, Liebman, and Katz 2007).¹⁵ This entails applying unit weights to the standardized z_j (Bobko, Roth, and Buster 2007). The index is the equally-weighted average of the standardized indicators, given by

$$x^z = \frac{1}{J} \sum_{j=1}^J \left(\frac{z_j - \bar{z}_j}{\sigma_j} \right) = \sum_{j=1}^J \left(\frac{1}{J\sigma_j} \right) z_j - \sum_{j=1}^J \left(\frac{1}{J\sigma_j} \right) \bar{z}_j = \zeta + \sum_{j=1}^J \lambda_j z_j, \quad (14)$$

¹² Bobko, Roth, and Buster (2007) provides an excellent overview.

¹³ Two well-known applications of the equally-weighted index are Solon (1992) and Ashenfelter and Krueger (1994). The former uses income averaged over an increasing number of years to proxy for permanent income and the latter considers averaging reports of individual education obtained by a respondent and their twin.

¹⁴ To be clear, the weights are fixed *ex ante*, so a leverage point in one variable cannot distort the weighting scheme, unlike PCA where the weights are data-driven. Of course, extreme z_j values still affect the index linearly, so the robustness claim pertains only to the choice of weights, not in the score itself.

¹⁵ A search on Google scholar for “‘mean z -score’ AND ‘index’ AND ‘regression’” produces more than 6,400 hits. Accessed on 01 April 2025.

where $\lambda_j = 1/J\sigma_j$ and ζ is a constant equal to $-\sum_j \lambda_j \bar{z}_j$. Thus, the mean z -score index is a linear combination of the indicators with weights that are decreasing in the variance of z_j and are not constrained to sum to one.

In the classical setting and analogous to $\text{plim} [\hat{\beta}^{\text{PCA}}]$, one can apply Equation (7) to obtain the probability limit of the OLS estimator $\hat{\beta}^z$:

$$\text{plim} [\hat{\beta}^z] = \beta \left[1 - \frac{1 + \left(1 - 2 \sum_j \lambda_j\right) \rho^z}{1 - R_{x^{\text{PCA}}, \mathbf{w}}^2} \right], \quad (15)$$

where $\rho^z := \text{Var}(x^*) / \text{Var}(x^z)$. As with PCA, the mean z -score index does not use the normalization that the weights sum to unity.¹⁶ Moreover, the effect of ζ in x^z vanishes as $N \rightarrow \infty$, implying that the behavior of $\hat{\beta}^z$ is identical to $\hat{\beta}^{\text{PCA}}$ except with different weights. Specifically, $\hat{\beta}^z$ does not necessarily suffer from attenuation bias even in the presence of classical measurement error in the individual indicators. It suffers from expansion bias if $\sum_j \lambda_j > (1 - \rho^z) / 2\rho^z$, which is guaranteed to be the case if $\rho^z = 1$. As the reliability ratio goes to zero, $\hat{\beta}^z$ will suffer from attenuation bias and may even flip the sign. In addition, $\hat{\beta}^z$ is consistent only if the numerator in Equation (15) is zero, which requires that $\rho^z = 1 / (2 \sum_j \lambda_j - 1)$. In the two-indicator example (Section 3.2), the weights on the raw variables using this approach would correspond to $\lambda_1 \approx 0.41$ and $\lambda_2 \approx 0.22$, yielding a reliability ratio of $\rho^z \approx 0.59$.

Partial Least Squares. Partial least squares (PLS) offers a “supervised” route to dimensionality reduction that remains rare in economics but is well established elsewhere. Developed by Wold in the 1960s (see, e.g., Wold 1982), PLS is now widely used in chemistry, psychometrics, and the life sciences.¹⁷ Like principal component analysis, PLS constructs orthogonal linear combinations of the observed variables $\mathbf{z} = (z_1, z_2, \dots, z_J)$. Unlike PCA, however, it is *supervised*: the response variable y explicitly guides the choice of weights. Formally, the first PLS component $x^{\text{PLS}} = \sum_j v_j z_j$ is obtained by selecting the unit-length vector \mathbf{v} that maximizes $\text{Cov}(x^{\text{PLS}}, y)$. PCA, by contrast, ignores y and instead maximizes $\text{Var}(x^{\text{PCA}})$. Subsequent PLS components are derived by (i) deflating both y and each z_j with respect to the preceding component and (ii) repeating the same covariance-maximization step on the resulting residuals. This recursive *orthogonal-scores* algorithm yields indices that are mutually uncorrelated and jointly maximize the explained covariance with y , which is arguably better in prediction problems (e.g., Geladi and Kowalski 1986).

Operationally, the first PLS component is obtained by solving

$$\mathbf{v}^{\text{PLS}} = \arg \max_{\|\mathbf{v}\|=1} \text{Cov}(\mathbf{v}'\mathbf{z}, y). \quad (16)$$

When the predictors are standardized beforehand, the problem is equivalent to maximizing the correlation between $x^{\text{PLS}} = \mathbf{v}^{\text{PLS}'}\mathbf{z}$ and y , ensuring scale invariance without additional constraints. With standardized

¹⁶ In fact, conditional on the sum of the mean z -score weights, the mean z -score index is suboptimal. In other words, if one were to choose weights constrained to sum to the same total to maximize the reliability ratio, the solution is not the weights used by the index.

¹⁷ A search on Google scholar for “‘partial least squares’ AND ‘index’ AND ‘regression’” produces more than 153,000 hits; 17,600 since 2024. Accessed on 16 April 2025.

\mathbf{z} , the solution has a closed-form expression:

$$v_j^{\text{PLS}} = \frac{\text{Cov}(z_j, y)}{\sqrt{\sum_{k=1}^J \text{Cov}(z_k, y)^2}}, \quad j = 1, 2, \dots, J. \quad (17)$$

Hence PLS weights are proportional to the covariances (not variances) of the individual indicators with the outcome. This distinction matters when the z_j are noisy indicators for an unobserved construct x^* . Under nondifferential measurement error— $\Pr(z_j | x^*, y) = \Pr(z_j | x^*)$ —we have $\text{Cov}(z_j, y) = \text{Cov}(x^*, y)$ for every j , so that all indicators receive identical weights after normalization. In expectation, PLS therefore collapses to an equal-weights index up to scale, with the only difference being that PLS enforces a unit-length constraint ($\|\mathbf{v}\| = 1$) whereas the equal-weights index imposes $\sum_j v_j = 1$, although this symmetry may break in finite samples.

Although data-adaptive, the PLS weights are not generally proportional to the inverse measurement-error variances and thus do not coincide with the optimal weights derived in Section 2.2. As a result, the probability limit of the OLS estimator that uses x^{PLS} as a regressor equals that in Equation (13), with the PCA weights and reliability factor replaced by their PLS counterparts, and therefore, subject to the same complex bias properties if $\sum_j v_j^{\text{PLS}} \neq 1$.

Exploratory Factor Analysis. Exploratory factor analysis (EFA) is the third approach consider for dimensionality reduction. Dijkstra (2010) contends that factor models are the most commonly used method in the social sciences. EFA proceeds by identifying the shared variance shared among the indicators and attributing 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, maximum likelihood, or alpha factoring) yield different estimates of the weights. However, a general insight holds: indicators with larger unique variances (i.e., more noise) receive smaller loadings. Equivalently, under classical measurement error, the implied weight is proportional to its signal-to-noise ratio, in contrast to PCA, which uses only the total variance. If the disturbances are correlated, some of that spurious covariance will be absorbed into the factors, so EFA can likewise be biased by measurement-error correlations. PCA suffers when error variances are large; EFA suffers when error correlations are nonzero. PLS is similar to EFA in its focus on common variance except PLS incorporates the outcome into the factor extraction process. In line with PCA and PLS, we focus only on the first factor.

4.2 Foregoing Index Creation in Regression (Lubotsky–Wittenberg Method)

Despite the reliance of PCA, the mean z -score index, and the equal index on sub-optimal weights and the fact that the optimal weights are unknown to the researcher, it is nonetheless possible to obtain an estimate, $\hat{\beta}^*$, based on the optimal weights. As Lubotsky and Wittenberg (2006) show, OLS estimation of

$$y = \alpha + \sum_j \beta_j z_j + \mathbf{w}'\boldsymbol{\gamma} + \varepsilon \quad (18)$$

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

$$y = \alpha + \beta x^* + \mathbf{w}'\boldsymbol{\gamma} + \varepsilon \quad (19)$$

where $x^* = \sum_j \lambda_j^* z_j$ is the optimal index.

In the case of two indicators, for example, the model is

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \mathbf{w}'\boldsymbol{\gamma} + \varepsilon. \quad (20)$$

Under the assumption that z_1 and z_2 are both measures of the same x^* (the classical measurement-error setting), Lubotsky and Wittenberg (2006) show that the OLS estimates $\widehat{\beta}_1^{\text{OLS}}$ and $\widehat{\beta}_2^{\text{OLS}}$ have a particular interpretation: the sum $\widehat{\beta}_1^{\text{OLS}} + \widehat{\beta}_2^{\text{OLS}}$ is algebraically identical to the estimate $\widehat{\beta}^{\text{OLS}}$ obtained from regression y on $x^* = \lambda_1^* z_1 + \lambda_2^* z_2$. In other words, by including multiple indicators, OLS automatically finds the best linear combination (in terms of minimizing bias) and estimates β accordingly.

This remarkable result, which generalizes to J manifest variables and correlation among the measurement errors u_j , suggests that one need not actually compute an index at all: one can run a regression with all indicators and then report the sum of their coefficients as the estimated effect of the latent factor. This procedure effectively achieves what PCA and other aggregation schemes such as the mean z -score index and the equal index fail to do: it optimally weights the manifest variables by their informativeness in explaining y . Importantly, it does so without requiring the researcher to know the indicators' noise variances *a priori*. The OLS estimation uses sample correlations to implicitly determine the weighting. Intuitively, the regression will automatically assign weights to each indicator via their estimated coefficients in proportion to the signal each indicator contains about the latent factor.

To the extent that one's goal is estimating β , this multi-indicator regression approach is an extremely attractive solution when feasible. It is simple to implement and does not require additional instruments, complex models, or ad hoc aggregation schemes. It also uses all available information: no potentially useful indicator is discarded or down-weighted arbitrarily. Furthermore, classical measurement-error analysis indicates that including more indicators increases the signal-to-noise ratio. In fact, if one had a large number of independent indicators, including them all would make the combined estimate of β approach the true β (i.e., the attenuation bias vanishes as $J \rightarrow \infty$). This result has seemingly gone unnoticed by applied researchers despite Lubotsky and Wittenberg (2006, p. 549) warning that the prevailing practice of creating summary measures such as PCA is “generally ad hoc and hardly ever optimal”.¹⁸

4.3 Optimal Index Construction via Nonlinear Regression (Yang–Jia–Li Method)

A straightforward extension to the Lubotsky and Wittenberg (2006) approach follows from Yang, Jia, and Li (2023). In Yang, Jia, and Li (2023), the authors consider the problem of aggregating from high frequency data to a lower frequency. For example, one may wish to aggregate monthly unemployment rates to the annual unemployment rate. While this issue stems from a mismatch between the frequency of the observed data and the researcher's desired, unobserved frequency rather than measurement error,

¹⁸ Lubotsky and Wittenberg (2006) has only 216 citations on Google Scholar as of 05 April 2025.

the way to proceed is identical.

Substitution of the definition of x^* into Equation (19) and using the normalization that $\sum_j \lambda_j = 1$, the estimating equation becomes

$$y = \alpha + \beta \sum_{j \neq J} \lambda_j z_j + \beta \left(1 - \sum_{j \neq J} \lambda_j \right) z_J + \mathbf{w}' \boldsymbol{\gamma} + \varepsilon. \quad (21)$$

While this is nonlinear in the parameters $\{\alpha, \beta, \boldsymbol{\lambda}, \boldsymbol{\gamma}\}$, the parameters are identified under the usual conditions and can be estimated using any nonlinear regression technique such as Generalized Method of Moments (GMM), Nonlinear Least Squares (NLS), or maximum likelihood. In fact, the parameter estimates are identical the Lubotsky and Wittenberg (2006) approach. However, the fact that the weights, $\boldsymbol{\lambda}$, are estimated along with α , β , and $\boldsymbol{\gamma}$, opens up two possibilities. First, the optimal index can be generated after-the-fact if desired. Second, the optimal weights can be compared to the weights chosen by PCA.

4.4 Instrumental Variables

As mentioned previously, it is well known that β can be consistently estimated by IV using one or more indicators to instrument for another indicator if the measurement errors are uncorrelated across indicators; more efficient estimates can be obtained by combining multiple IV estimates (e.g., Andersson and Møen 2016; Gillen, Snowberg, and Yariv 2019). However, this is often difficult to justify in practice, particularly with bounded indicators (Black, Berger, and Scott 2000).¹⁹ If the independence assumption is violated, the probability limit of the IV estimator is

$$\text{plim } \widehat{\beta}^{\text{IV}} = \beta \left[\frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Cov}(u_{j'}, u_j)} \right]. \quad (22)$$

in the regression model with no other covariates.²⁰ The IV estimate will suffer from attenuation bias if $\text{Cov}(u_{j'}, u_j) > 0$ and expansion bias otherwise. However, even if IV is consistent, it does not necessarily dominate the optimal index approach of Lubotsky and Wittenberg (2006) or Yang, Jia, and Li (2023) in terms of mean squared error, particularly if the correlation between the indicators is weak or in the presence of high leverage observations and clustered and heteroskedastic errors, due to the relative imprecision of the IV estimator (e.g., Young 2022).

When the indicators are correlated and ‘traditional’ IV using one or more indicators to instrument for a different indicator is inconsistent, it is not well known that several empirical methods have been developed recently to address situations with invalid (so-called ‘imperfect’) instruments. Examples include Conley, Hansen, and Rossi (2012), Kippersluis and Rietveld (2018), and Chalak and Kim (2024), where the instrument is allowed to directly enter the structural model for y and Nevo and Rosen (2012), where the instrument is allowed to be correlated with the error in the structural model. Alternatively, one may

¹⁹ In a canonical example, Ashenfelter and Krueger (1994) consider this issue in the context of using a twin’s report of their sibling’s education as an instrument the sibling’s self-reported education.

²⁰ In a multiple regression, Equation (22) still holds but with x^* , u_j , and $u_{j'}$ replaced with their residuals after removing \mathbf{w} following the Frisch-Waugh-Lovell theorem.

exploit higher moments of the index for identification as in Klein and Vella (2010), Lewbel (2012), and Lewbel, Schennach, and Zhang (2024). Rather than focusing on estimation, Kim and Wilhelm (2024) concentrate on inference and propose a more powerful t -test to test the null that β equals zero in the situation with two indicators with correlated measurement errors. The procedure searches over all linear combinations to include as the index and all other linear combinations to use as the instrument. Finally, other papers such as Ashley and Parmeter (2015), Kiviet (2016), Kiviet (2020), Kripfganz and Kiviet (2021), and Kiviet (2023) assess what can be learned from the OLS estimates in the presence of endogeneity and no valid instruments. For brevity, we do not examine these approaches here as there is nothing unique about their performance in the current situation.

5 Alternative Approaches When the Index is a Nuisance Function

In situations where the effect of the index, β , is not of interest, but rather the focus is on γ , then the unknown index, $g(\mathbf{z})$, is a nuisance function and β is a nuisance parameter. In this case, we consider three estimators to remove or control for the index. To proceed, we re-write Equation (1) as

$$y = \alpha + g(\mathbf{z}) + \mathbf{w}'\gamma + \varepsilon. \quad (23)$$

Our first estimator approximates the unknown $g(\mathbf{z})$ using a series approach and then estimates the model using Lasso. This is similar to Ash, Galletta, and Giommoni (2025) who use other machine learning methods to predict and create an index of corruption. Here, the Lasso estimator minimizes the objective function given by

$$\min_{\alpha, \gamma, \kappa} \sum_i \varepsilon_i^2 + \psi ||\kappa||, \quad (24)$$

where ψ is a tuning parameter and κ are coefficients in the series approximation given by

$$g(\mathbf{z}) \approx \ddot{\mathbf{z}}' \kappa \quad (25)$$

and $\ddot{\mathbf{z}}$ is a vector containing a p^{th} -order polynomial in the elements of \mathbf{z} . In practice, we set $p = 3$ and choose ψ via K -fold cross validation with $K = 5$.

Our second and third estimators build on the logic of matching estimators. Specifically, if for each observation i we can find another observation j with identical \mathbf{z} , then differencing the observations removes the nuisance function, $g(\mathbf{z})$ from the estimating equation. Formally,

$$y_i - y_j = (\mathbf{w}_i - \mathbf{w}_j)'\gamma + \varepsilon_i - \varepsilon_j. \quad (26)$$

As with matching in other contexts, if \mathbf{z} is high dimensional and/or includes continuous covariate(s), then *exact* matching is not feasible. While there are many ways to implement *inexact* matching (or *coarsened exact* matching), we compute the distance between \mathbf{z}_i and \mathbf{z}_j using the Mahalanobis metric. We then find the M closest matches and subtract the average of these matches. Denote the set of matches for

observation i as \mathcal{M}_i . Thus, our second estimator uses OLS to estimate

$$y_i - \tilde{y}_i = (\mathbf{w}_i - \tilde{\mathbf{w}}_i)' \boldsymbol{\gamma} + \varepsilon_i - \tilde{\varepsilon}_i. \quad (27)$$

where $\tilde{q}_i := (1/\#\mathcal{M}_i) \sum_{j \in \mathcal{M}_i} q_j$ for $q = \{y, \mathbf{w}, \varepsilon\}$. Our third estimator combines matching with regression adjustment to remove residual differences in \mathbf{z} that remain due to the inexactness of the matching. We use Lasso to estimate

$$y_i - \tilde{y}_i = \ddot{\mathbf{z}}_i' \boldsymbol{\kappa}_1 + \ddot{\mathbf{z}}_i' \boldsymbol{\kappa}_2 + (\mathbf{w}_i - \tilde{\mathbf{w}}_i)' \boldsymbol{\gamma} + \varepsilon_i - \tilde{\varepsilon}_i, \quad (28)$$

where $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ are penalized. In practice, we set $M = 10$ and continue to set $p = 3$.

When the data-generating process is the formative indicators model in Figure 1 and the full set of indicators is known, then all three estimators should perform well in terms of estimating $\boldsymbol{\gamma}$ as the Lasso and/or matching precludes the need to specify the functional form of the index. Under the reflective indicators model, or the formative indicators model where not all indicators are observed, then the estimates of $\boldsymbol{\gamma}$ will likely be severely biased.

6 Simulation-Based Comparison

The simulation is designed to mimic a common scenario: an unobserved variable x^* influences an outcome y , and researchers have multiple indicators for x^* . We will examine how different index construction strategies fare in estimating the effect of both x^* and \mathbf{w} on y . We specify the true β and $\boldsymbol{\gamma}$ and measure bias, variance, and mean squared error of each estimator across many repeated samples.

6.1 Simulation Design

We assess five experimental designs, each nested in the following data-generating process (DGP):

$$y_i = \alpha + \beta x_i^* + \gamma w_i + \varepsilon_i, \quad i = 1, 2, \dots, N; \quad (29)$$

$$z_{ji} = \omega_j x_{ji}^* + u_{ji}, \quad j = 1, 2, \dots, J. \quad (30)$$

which is very similar to that used in Andersson and Møen (2016). Across all experiments, we set $\alpha = 0$, $\beta = 1$, and $N = 10,000$ and draw $x^*, w \stackrel{\text{iid}}{\sim} \mathcal{N}_2(0, 0, 1, 1, 0.5)$ and $z_1, \dots, z_J \stackrel{\text{iid}}{\sim} \mathcal{N}_J(0, \Sigma_u)$, where \mathcal{N}_k is a k -dimensional multivariate normal distribution. We iterate these data generation and estimation processes for 1,000 repetitions to obtain the sampling distribution of each considered estimator. Our evaluation focuses on the estimated slope coefficient(s) and hinges on three criteria: (1) the mean bias, (2) the root mean squared error (RMSE) of the estimator, serving as a measure of its overall accuracy and efficiency, and (3) the empirical coverage probability of the nominal 95% confidence interval(s).

We compare the performance of several estimators: (1) OLS using the true latent variable x^* as a benchmark, (2) OLS including each indicator individually in the regression, (3) OLS including an index obtained using the first principal component derived from the J indicators after standardization, (4) OLS including an equally-weighted average of the indicators in their original form (equally-weighted

index) and standardized versions (mean z -score index), (5) OLS including an index obtained using the first component from partial least squares derived from the J indicators after standardization, (6) OLS including an index obtained using the first factor score from three different exploratory factor analysis techniques derived from the J indicators after standardization, (7) OLS including all J indicators as covariates and summing the coefficients (Lubotsky and Wittenberg (2006) approach), (8) GMM including all J indicators as covariates and estimating the optimal index weights—where the weights are restricted to sum to unity but may or may not be restricted to the unit interval—along with the slope coefficient(s) (Yang, Jia, and Li (2023) approach), and (9) IV (estimated via two-stage least squares or 2SLS) using each combination of $J - 1$ indicators to instrument for the remaining indicator.²¹ In addition, we record and graph the PCA, PLS, and EFA loadings and GMM optimal index weights across simulation replications to assess how differently these methods combine the indicators to approximate x^* .

The five experimental designs are as follows.

- SCENARIO 1: $\gamma = 0$, $J = 3$, $\lambda_j = 1 \forall j$, and the covariance matrix for the measurement errors is given by

$$\Sigma_{u,1} = \begin{bmatrix} 0.4 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 4.0 \end{bmatrix}.$$

Here, each manifest variable z_j is a noisy indicator of the latent variable x^* with different noise levels. These values correspond to indicators with reliability ratio $\rho_j = \text{Var}(x^*) / \text{Var}(z_j)$ of approximately 0.70, 0.50, and 0.20, respectively. Indicator 1, z_1 , is high-quality; z_2 is medium-quality; and the last indicator, z_3 , is quite noisy. This spread reflects a situation where one indicator is fairly closely tied to the latent concept while others are very noisy. We include no other covariates \mathbf{w} in this simulation for simplicity, focusing on estimating β .

- SCENARIO 2: $\gamma = 0.5$, $J = 3$, $\lambda_j = 1 \forall j$, and the covariance matrix for the measurement errors is given by

$$\Sigma_{u,2} = \begin{bmatrix} 0.4 & 0.3 & 0.2 \\ 0.3 & 1.0 & 0.4 \\ 0.2 & 0.4 & 4.0 \end{bmatrix}.$$

Here, the measurement errors are correlated. We also include a second covariate, w , that is observed and has a correlation of 0.5 with x^* . For example, Radatz and Baten (2025) explores the impact of an index of inequality, which is a weighted average of Gini coefficients for income, height, and land, on civil conflict. The authors also control for other factors such as an index for institutional quality and ethnic fractionalization, among others.

- SCENARIO 3: $\gamma = 0.5$, $J = 3$, $\lambda_j = 1 \forall j$, and the covariance matrix for the measurement errors is given by

$$\Sigma_{u,3} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.4 \\ 0.0 & 0.4 & 4.0 \end{bmatrix}.$$

²¹ Simulations are performed in Stata 18 using `pca`, `pls`, `factor`, `regress`, `gmm`, and `ivregress 2sls`.

Here, the first indicator is perfect in the sense that it is measured without error (i.e., $z_1 = x^*$), while the remaining indicators continue to be noisy with correlated measurement errors. We assume the researcher does not know this *ex ante*. This scenario allows us to investigate how the various estimation methods behave in the (unknown) presence of a perfect indicator.

- SCENARIO 4: $\gamma = 0.5$, $J = 3$, $\lambda_1 = 1$, $\lambda_2 \in \{1.5, 5, 10\}$, $\lambda_3 = 0.5$, and the covariance matrix for the measurement errors is given by $\Sigma_{u,2}$. Here, the second and third indicators are of a different scale than the latent x^* . For example, Montero and Yang (2022) form an index of municipality-level development where indicators are measured in people (population), currency (income), years (education), and percentages (e.g., employment rate). Akee et al. (2018) construct an index of ‘agreeableness’ where indicators are counts (e.g., number of arguments), binary indicators (e.g., cruelty to animals), among others. Maccini and Yang (2009) constructs an asset index where one indicator is in currency (asset values) and others are binary indicators (e.g., own a television). In addition to the estimators listed previously, we also consider an alternative estimator proposed in Lubotsky and Wittenberg (2006) where $\hat{\beta}$ is the weighted sum of the OLS coefficients on the individual indicators, where the weight on z_1 is one and the weight on z_j , $j = 2, 3$, is $\text{Cov}(z_j, y) / \text{Cov}(z_1, y)$.
- SCENARIO 5: $\gamma = 0.5$, $J \in \{3, 5, 10, 15, 20, 30, 40, 50\}$, and the indicators are generated as

$$z_{ji} = x_i^* + 0.5\eta_i^{(1)} + \frac{J}{3}\eta_{ji}^{(2)},$$

where $\eta^{(1)}, \eta_1^{(2)}, \dots, \eta_J^{(2)} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Thus, the measurement error is correlated across all indicators and the additional indicators are increasingly noisy. This scenario allows us to investigate how incorporating additional indicators containing little signal influences the different estimators. For example, Blattman, Fiala, and Martinez (2013) construct an index of individual income/consumption using 70 indicators and Montero and Yang (2022) form an index of municipality-level development based on 35 indicators. Stoetzer, Zhou, and Steenbergen (2025) finds that efficiency of there latent average treatment effect estimator improves as the number of reliable indicators increases. We do not apply all estimators in this scenario for simplicity. We focus on the benchmark case, PCA, the equally-weighted index, the mean z -score index, and GMM without restricting the weights to be in the unit interval.

The preceding designs model the indicators, z_j , $j = 1, \dots, J$ as error-ridden indicators for the latent construct, x^* . This takes the latent construct as fixed and expresses the indicators as functions of the construct. This presupposes a particular way of conceptualizing the relationship between the indicators and the latent construct, where the arrows in a causal graph go from the latent construct to the indicators. Stoetzer, Zhou, and Steenbergen (2025) refer to this set up as a *reflective indicators model* as the indicators reflect the latent construct. This conceptualization may be especially unfavorable to PCA. As an alternative, referred to as a *formative indicators model*, the indicators are inputs into the latent construct. Here, the arrows in a causal graph go from the indicators to the latent construct. For example, the z ’s may represent various aspects of individual health and x^* is latent aggregate of these indicators.

Or, the z 's may capture particular dimensions of freedom (e.g., speech, press, elections, etc.) and x^* is the latent aggregate representing overall freedom. From this perspective, x^* is still latent due to not observing all inputs and not knowing the 'true' aggregation scheme. Nonetheless, the z 's are not viewed as being error-laden. PCA may fare better in this setup since maximizing the variation in x may lead to a less flawed econometric specification.

To proceed, we assess an additional experimental design according to the following DGP.

- SCENARIO 6:

$$y_i = \alpha + \beta x_i^* + \gamma w_i + \varepsilon_i, \quad i = 1, 2, \dots, N \quad (31)$$

$$w_i = c_i + u_{wi} \quad (32)$$

$$z_{ji} = c_i + u_{ji}, \quad j = 1, 2, \dots, \mathcal{J} \quad (33)$$

$$x^* = \sum_{j=1}^{\mathcal{J}} \delta_j^* z_j^* \quad (34)$$

where $c, u_w, u_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $\delta_j \stackrel{\text{iid}}{\sim} \mathcal{U}[0, 1]$, $\delta_j^* = \delta_j / \sum_{j'=1}^{\mathcal{J}} \delta_{j'}$, and z_j^* is the standardized version of z . This design allows for correlation between w and x^* and for x^* to be a weighted average of the full set of \mathcal{J} indicators, z_j , where the weights sum to one but are randomized so that the importance of the individual indicators varies. We assume that the researcher only observes the first $J < \mathcal{J}$ indicators. We set $J = 3$ and consider $\mathcal{J} \in \{5, 20\}$. When $\mathcal{J} = 5$ (20), the majority of the indicators is (is not) observed by the researcher.

6.2 Simulation Results

Table 1 reports the performance of various estimators under the baseline scenario of uncorrelated measurement errors, with the methods sorted in order of increasing RMSE. As expected, using the true latent variable, x^* , recovers the true parameter $\beta = 1$ without bias (RMSE = 0.010), thereby serving as a benchmark for the other methods.

The instrumental variables (IV) estimators—implemented with three different instrument sets (using (z_2, z_3) as instruments for z_1 , (z_1, z_2) for z_3 , and (z_1, z_3) for z_2)—yield average estimates that are essentially unbiased (ranging from 0.999 to 1.000) with excellent coverage probabilities (approximately 94–96%). This unbiasedness, however, comes at the cost of slightly increased dispersion (RMSE values ranging from 0.016 to 0.025), reflecting the common bias–efficiency tradeoff observed in IV estimation.

Using the mean z -score index produces an average estimate of 1.073, reflecting an upward bias of approximately 7.3%. Its RMSE is correspondingly increased to 0.074. In the multi-indicator regression (the Lubotsky and Wittenberg (2006) approach), which aggregates information from all indicators directly, the average estimate is 0.781, corresponding to an attenuation bias of approximately 22%, with an RMSE of 0.219. The coverage probabilities indicate that the confidence intervals substantially underestimate true parameter uncertainty.

Using individual indicators separately reveals even more pronounced discrepancies. The best-performing single indicator (z_1) yields an average estimate of 0.699, a 30% attenuation bias, and an

TABLE 1
SCENARIO 1: ALL INDICATORS NOISY WITH UNCORRELATED ERRORS

Method	$\hat{\beta}$	Bias	Coverage	$\sigma_{\hat{\beta}}$	RMSE
True x^*	1.000	0.000	0.937	0.010	0.010
Factor Index 3	0.999	-0.001	0.920	0.014	0.014
IV ($z_2, z_3 \rightarrow z_1$)	0.999	-0.001	0.938	0.016	0.016
IV ($z_1, z_3 \rightarrow z_2$)	1.000	0.000	0.956	0.017	0.017
IV ($z_1, z_2 \rightarrow z_3$)	0.999	-0.001	0.954	0.025	0.025
Mean z -Score Index	1.074	0.074	0.001	0.015	0.075
Factor Index 1	1.087	0.087	0.000	0.015	0.088
PLS Index	0.874	-0.126	0.000	0.013	0.127
Factor Index 2	0.859	-0.141	0.000	0.013	0.141
GMM	0.789	-0.211	0.000	0.010	0.211
All Indicators	0.789	-0.211	0.000	0.010	0.211
Single Indicator z_1	0.714	-0.286	0.000	0.010	0.286
PCA Index	0.628	-0.372	0.000	0.009	0.372
Equal-weight Index	0.625	-0.375	0.000	0.009	0.375
Single Indicator z_2	0.499	-0.501	0.000	0.009	0.501
Single Indicator z_3	0.200	-0.800	0.000	0.006	0.800

NOTES.—The table presents the performance of various estimation methods for β in the model $y = \alpha + \beta x^* + \varepsilon$, where x^* is unobserved. Data were simulated with 10,000 observations and 1,000 repetitions. Three indicators (z_1 , z_2 , and z_3) for x^* were generated with uncorrelated errors. The true value of β is 1, and $\alpha = 0$. The table shows the average estimated β ($\hat{\beta}$), bias, coverage probability of 95% confidence intervals (Coverage), standard deviation of $\hat{\beta}$ ($\sigma_{\hat{\beta}}$), and root mean squared error (RMSE) for each method. The IV rows report 2SLS estimates using two indicators as instruments for the third. The GMM row uses all three indicators in a generalized method of moments framework. Indices (PCA, PLS, equal-weight, and mean z -score) combine all indicators.

RMSE of 0.301. This performance is identical to that of the generalized method of moments (GMM) estimator (also 0.781, bias = -0.219 , RMSE = 0.219). The remaining indicators exhibit considerably worse performance: z_2 produces an estimate of 0.499 (a 50% bias), while the poorest indicator, z_3 , achieves an estimate of merely 0.200 (an 80% bias). These outcomes align with their respective reliability levels by construction. The PCA and equal-weight indices yield average estimates of 0.627 and 0.623, respectively, both reflecting biases exceeding 37%, underscoring that standard dimension-reduction methods such as PCA do not optimally utilize indicator information when some indicators are highly noisy.

In this baseline scenario with uncorrelated measurement errors, our simulation results reinforce the well-known attenuation bias inherent in using imperfect indicators. While IV methods demonstrate robustness in recovering the true effect of x^* on y , alternative approaches—including the mean z -score index, multi-indicator regression, GMM, single-indicator regressions, and conventional indices—display biases (often substantial) and significant inaccuracies in coverage probabilities, further emphasizing the importance of careful methodological selection in empirical analyses involving latent constructs.

Tables 2 and 6 present detailed results from the simulation exercises aimed at evaluating the performance of several estimation strategies when indicators for a latent variable are available but measured with error. To reiterate, in our simulation framework, we specify the true model parameters as $\beta = 1$, $\gamma = 0.5$, and $\alpha = 0$. We generate synthetic datasets of 10,000 observations and repeat the simulation 1,000 times to obtain robust empirical sampling distributions for each estimator.

Table 2 summarizes outcomes under Scenario 1, where all three indicators are noisy indicators of the latent construct, with correlated measurement errors. This scenario reflects a realistic challenge often encountered in applied econometric analysis. Across all evaluated estimation methods, we observe significant bias in estimates of β and all but the true x^* have a non-zero coverage probability. This underscores the inherent difficulty in consistently estimating regression parameters when no available indicator perfectly measures the latent construct. Specifically, methods such as the equal-weight index and PCA index produce notable attenuation biases (-0.516 and -0.506 , respectively), highlighting a fundamental limitation of these common approaches: their reliance on maximizing explained variance rather than explicitly minimizing measurement error.

Instrumental-variable (IV) approaches, which attempt to leverage relationships between indicators strategically, still demonstrate substantial biases (ranging from approximately -0.27 to -0.29). These IV estimates, while somewhat improved compared to simpler indices, nonetheless reveal persistent difficulties in achieving unbiased estimation in the presence of correlated measurement errors. Similarly, the generalized method of moments (GMM) approach and the Lubotsky and Wittenberg (2006) strategy of including all indicators directly in regression—numerically equivalent in this case—also yield significant attenuation bias, around -0.336 . While these two estimators are insufficient to address the measurement problem, they are demonstrably superior to using the PCA index. Among individual indicators examined separately, even the best-performing indicator (z_1) still exhibits considerable attenuation bias (-0.347), while the noisiest indicator (z_3) yields severe attenuation bias (-0.842).

In contrast, Table 6 presents results under Scenario 2, where the first indicator (z_1) is perfectly measured without error, although researchers typically lack prior knowledge of this advantageous property.

TABLE 2
SCENARIO 2: ALL INDICATORS NOISY WITH CORRELATED ERRORS

Method	$\hat{\beta}$	Bias (β)	Coverage (β)	$\hat{\gamma}$	Bias (γ)	Coverage (γ)	$\sigma_{\hat{\beta}}$	$\sigma_{\hat{\gamma}}$	RMSE (β)	RMSE (γ)
True x^*	1.001	0.001	0.935	0.499	-0.001	0.946	0.012	0.012	0.012	0.012
Mean z-score Index	0.842	-0.158	0.000	0.719	0.219	0.000	0.016	0.012	0.159	0.219
Factor Index 1	0.836	-0.164	0.000	0.691	0.191	0.000	0.015	0.012	0.165	0.191
Factor Index 3	0.804	-0.196	0.000	0.685	0.185	0.000	0.015	0.013	0.196	0.186
PLS Index	0.741	-0.259	0.000	0.695	0.195	0.000	0.014	0.012	0.259	0.195
IV ($z_1, z_2 \rightarrow z_3$)	0.726	-0.274	0.000	0.637	0.137	0.000	0.019	0.020	0.275	0.138
IV ($z_2, z_3 \rightarrow z_1$)	0.722	-0.278	0.000	0.638	0.138	0.000	0.014	0.013	0.278	0.139
Factor Index 2	0.722	-0.278	0.000	0.709	0.209	0.000	0.014	0.012	0.279	0.209
IV ($z_1, z_3 \rightarrow z_2$)	0.710	-0.290	0.000	0.645	0.145	0.000	0.013	0.014	0.291	0.145
GMM	0.664	-0.336	0.000	0.667	0.167	0.000	0.011	0.012	0.336	0.168
All Indicators	0.664	-0.336	0.000	0.667	0.167	0.000	0.011	0.012	0.336	0.168
Single Indicator z_1	0.653	-0.347	0.000	0.673	0.173	0.000	0.011	0.012	0.348	0.173
PCA Index	0.494	-0.506	0.000	0.709	0.209	0.000	0.009	0.012	0.506	0.209
Equal-weight Index	0.484	-0.516	0.000	0.757	0.257	0.000	0.010	0.012	0.516	0.258
Single Indicator z_2	0.429	-0.571	0.000	0.785	0.285	0.000	0.009	0.013	0.571	0.285
Single Indicator z_3	0.158	-0.842	0.000	0.920	0.420	0.000	0.006	0.013	0.842	0.420

NOTES.—The table presents the performance of various estimation methods for β and γ in the model $y = \alpha + \beta x^* + \gamma w + \varepsilon$, where x^* is unobserved. Data were simulated with 10,000 observations and 1,000 repetitions. Three indicators (z_1 , z_2 , and z_3) for x^* were generated with correlated errors with variance-covariance matrix described in the text. The true values are $\beta = 1$, $\gamma = 0.5$, and $\alpha = 0$. The table shows the average estimated coefficients, bias, coverage probability of 95% confidence intervals, standard deviation of the estimated coefficients, and root mean squared error (RMSE) for each parameter and method. The IV rows report 2SLS estimates using two indicators as instruments for the third. The GMM row uses all three indicators in a generalized method of moments framework. Indices (PCA, PLS, equal-weight, and mean z-score) combine all indicators.

In this scenario, methods explicitly designed to recognize or implicitly accommodate the presence of an error-free indicator deliver dramatically improved performance. Estimation strategies using the perfect indicator either alone or jointly with additional indicators achieve virtually unbiased estimates, with estimated coefficients consistently close to their true values and coverage probabilities approximating the nominal 95% confidence level. Notably, GMM and IV approaches excel in this scenario, demonstrating their capability to automatically assign greater weight to the high-quality indicator, yielding essentially unbiased estimates ($\hat{\beta}$ very close to 1) and reliably high coverage rates.

However, even under this favorable scenario, indices based on equal weighting or PCA continue to produce significant biases (-0.462 and -0.432 , respectively), revealing their structural limitations. Such methods, inherently designed to maximize variance rather than minimize measurement error, cannot distinguish between high- and low-quality indicators, leading to systematic misweighting and biased parameter estimates.

(Add graphs.)

Taken together, these detailed simulation findings emphasize the critical risks inherent in using PCA and other simple composite indices in regression analyses involving latent constructs. Across all simulations, the findings align with the analytical expectations that using the PCA index will consistently underestimate the relationship between the latent variable and the outcome of interest, often substantially. Our analysis strongly supports the adoption of estimation techniques explicitly constructed to manage measurement error, such as instrumental variables and GMM-based weighting schemes, as these methods consistently outperform simpler indexing approaches. Consequently, we recommend that empirical researchers exercise considerable caution when constructing composite indices from indicators and strongly advocate the selection of estimation methods that are explicitly structured to address and mitigate measurement errors to enhance the reliability and validity of inference and subsequent policy implications.

(Scenario 4.)

(Scenario 5.)

TABLE 3
SCENARIO 3: ONE INDICATOR IS PERFECT; REMAINING INDICATORS ARE NOISY

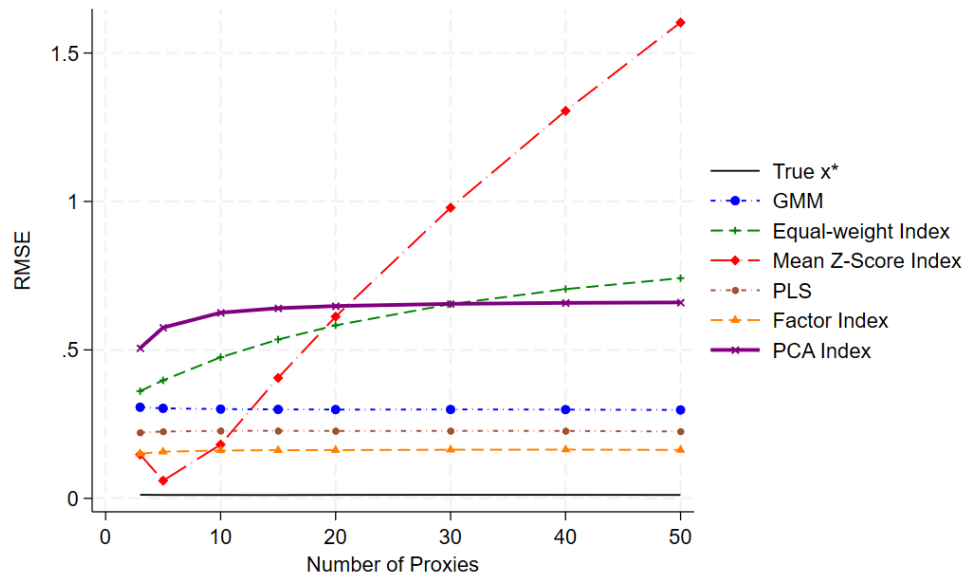
Method	$\hat{\beta}$	Bias (β)	Coverage (β)	$\hat{\gamma}$	Bias (γ)	Coverage (γ)	$\sigma_{\hat{\beta}}$	$\sigma_{\hat{\gamma}}$	RMSE (β)	RMSE (γ)
True x^*	1.000	0.000	0.955	0.500	0.000	0.948	0.011	0.012	0.011	0.012
Single Indicator z_1	1.000	0.000	0.955	0.500	0.000	0.948	0.011	0.012	0.011	0.012
All Indicators	1.000	0.000	0.955	0.500	0.000	0.949	0.011	0.012	0.011	0.012
GMM	1.000	0.000	0.954	0.500	0.000	0.948	0.011	0.012	0.011	0.012
GMM, [0,1] weights	0.999	-0.001	0.955	0.500	0.000	0.953	0.012	0.012	0.012	0.012
Factor Index 1	0.998	-0.002	0.000	0.606	0.106	0.000	0.015	0.013	0.015	0.107
IV ($z_2, z_3 \rightarrow z_1$)	0.999	-0.001	0.939	0.500	0.000	0.947	0.018	0.014	0.018	0.014
Mean z-score Index	0.963	-0.037	0.303	0.654	0.154	0.000	0.015	0.013	0.040	0.155
Factor Index 3	0.960	-0.040	0.000	0.594	0.094	0.000	0.018	0.015	0.044	0.095
IV ($z_1, z_3 \rightarrow z_2$)	0.949	-0.051	0.114	0.525	0.025	0.632	0.016	0.017	0.053	0.030
PLS Index	0.882	-0.118	0.000	0.592	0.092	0.000	0.012	0.012	0.118	0.093
IV ($z_1, z_2 \rightarrow z_3$)	0.825	-0.175	0.000	0.587	0.087	0.035	0.022	0.023	0.176	0.090
Factor Index 2	0.819	-0.181	0.000	0.640	0.140	0.000	0.013	0.013	0.182	0.140
PCA Index	0.568	-0.432	0.000	0.640	0.140	0.000	0.009	0.013	0.432	0.140
Equal-weight Index	0.538	-0.462	0.000	0.731	0.231	0.000	0.010	0.013	0.462	0.231
Single Indicator z_2	0.429	-0.571	0.000	0.786	0.286	0.000	0.009	0.013	0.572	0.286
Single Indicator z_3	0.158	-0.842	0.000	0.921	0.421	0.000	0.006	0.014	0.842	0.421

NOTES.—The table presents the performance of various estimation methods for β and γ in the model $y = \alpha + \beta x^* + \gamma w + \varepsilon$, where x^* is unobserved. The first indicator z_1 is measured without error ($z_1 = x^*$ exactly), while the remaining indicators (z_2 and z_3) are noisy. Data were simulated with 10,000 observations and 1,000 repetitions. The true values are $\beta = 1$, $\gamma = 0.5$, and $\alpha = 0$. The table shows the average estimated coefficients, bias, coverage probability of 95% confidence intervals, standard deviation of the estimates, and root mean squared error (RMSE) for each parameter and method. The IV rows report 2SLS estimates using two indicators as instruments for the third, and the GMM rows report estimates using a generalized method of moments framework. Indices (PCA, PLS, equal-weight, and mean z-score) combine all indicators.

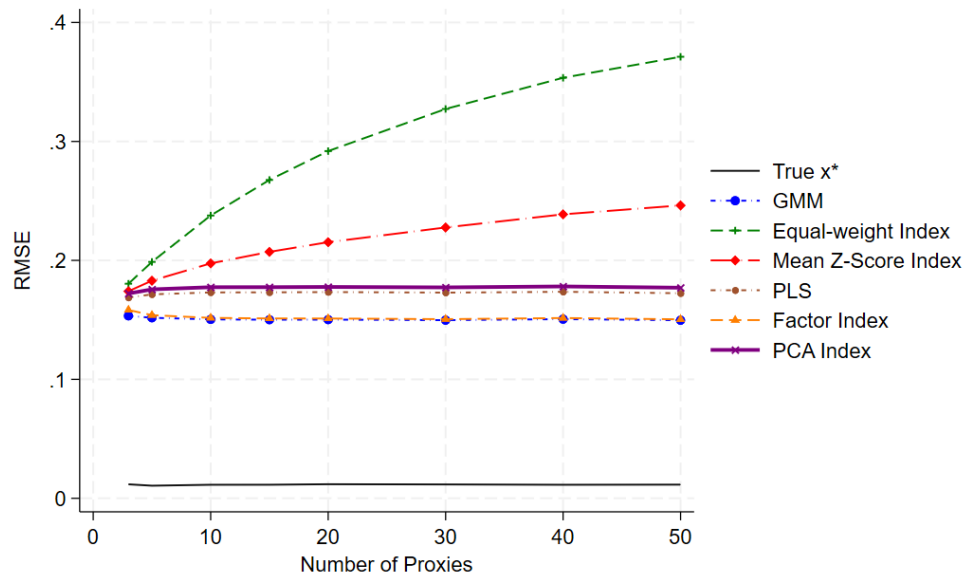
TABLE 4
SCENARIO 4: THREE INDICATORS ON DIFFERENT SCALES

Method	$\hat{\beta}$	Bias (β)	Coverage (β)	$\hat{\gamma}$	Bias (γ)	Coverage (γ)	$\sigma_{\hat{\beta}}$	$\sigma_{\hat{\gamma}}$	RMSE (β)	RMSE (γ)
<i>Panel A. $\lambda_2 = 1.5$</i>										
True x^*	1.001	0.001	0.935	0.499	-0.001	0.946	0.012	0.012	0.012	0.012
Factor Index 1	0.874	-0.126	0.000	0.649	0.149	0.000	0.015	0.012	0.127	0.150
Mean z -score Index	0.872	-0.128	0.000	0.721	0.221	0.000	0.016	0.012	0.129	0.221
Factor Index 3	0.822	-0.178	0.000	0.657	0.157	0.000	0.017	0.013	0.179	0.158
PLS Index	0.794	-0.206	0.000	0.658	0.158	0.000	0.014	0.012	0.207	0.158
IV ($z_1, z_2 \rightarrow z_3$)	1.205	0.205	0.000	0.698	0.198	0.000	0.044	0.028	0.210	0.200
IV ($z_2, z_3 \rightarrow z_1$)	0.788	-0.212	0.000	0.605	0.105	0.000	0.013	0.013	0.212	0.106
Factor Index 2	0.761	-0.239	0.000	0.681	0.181	0.000	0.014	0.012	0.239	0.182
GMM: rho weighted	0.709	-0.291	0.000	0.645	0.145	0.000	0.012	0.012	0.291	0.145
Single Indicator z_1	0.653	-0.347	0.000	0.673	0.173	0.000	0.011	0.012	0.348	0.173
GMM	0.594	-0.406	0.000	0.645	0.145	0.000	0.011	0.012	0.406	0.145
All Indicators	0.594	-0.406	0.000	0.645	0.145	0.000	0.011	0.012	0.406	0.145
PCA Index	0.536	-0.464	0.000	0.681	0.181	0.000	0.010	0.012	0.464	0.182
IV ($z_1, z_3 \rightarrow z_2$)	0.522	-0.478	0.000	0.608	0.108	0.000	0.009	0.013	0.478	0.109
Equal-weight Index	0.484	-0.516	0.000	0.757	0.257	0.000	0.010	0.012	0.516	0.258
Single Indicator z_2	0.419	-0.581	0.000	0.685	0.185	0.000	0.007	0.012	0.581	0.186
Single Indicator z_3	0.090	-0.910	0.000	0.977	0.477	0.000	0.006	0.013	0.910	0.477
<i>Panel B. $\lambda_2 = 5$</i>										
True x^*	1.000	0.000	0.955	0.500	0.000	0.948	0.011	0.012	0.011	0.012
Mean z -score Index	0.967	-0.033	0.429	0.667	0.167	0.000	0.015	0.012	0.036	0.167
Factor Index 1	0.962	-0.038	0.000	0.576	0.076	0.000	0.014	0.012	0.040	0.077
GMM: rho weighted	0.952	-0.048	0.030	0.524	0.024	0.447	0.014	0.012	0.050	0.027
IV ($z_2, z_3 \rightarrow z_1$)	0.921	-0.079	0.000	0.539	0.039	0.173	0.013	0.013	0.080	0.042
PLS Index	0.898	-0.102	0.000	0.580	0.080	0.000	0.013	0.012	0.103	0.081
Factor Index 3	0.897	-0.103	0.000	0.600	0.100	0.000	0.024	0.017	0.105	0.102
Factor Index 2	0.856	-0.144	0.000	0.613	0.113	0.000	0.013	0.012	0.145	0.113
Single Indicator z_1	0.652	-0.348	0.000	0.674	0.174	0.000	0.010	0.013	0.348	0.174
PCA Index	0.600	-0.400	0.000	0.613	0.113	0.000	0.009	0.012	0.400	0.113
IV ($z_1, z_2 \rightarrow z_3$)	1.423	0.423	0.000	0.645	0.145	0.014	0.057	0.033	0.427	0.149
Equal-weight Index	0.376	-0.624	0.000	0.593	0.093	0.000	0.005	0.012	0.624	0.093
Single Indicator z_2	0.190	-0.810	0.000	0.525	0.025	0.407	0.002	0.012	0.810	0.028
IV ($z_1, z_3 \rightarrow z_2$)	0.185	-0.815	0.000	0.538	0.038	0.157	0.003	0.013	0.815	0.040
GMM	0.134	-0.866	0.000	0.524	0.024	0.447	0.015	0.012	0.866	0.027
All Indicators	0.134	-0.866	0.000	0.524	0.024	0.442	0.015	0.012	0.866	0.027
Single Indicator z_3	0.090	-0.910	0.000	0.978	0.478	0.000	0.007	0.014	0.910	0.478
<i>Panel C. $\lambda_2 = 10$</i>										
True x^*	0.999	-0.001	0.949	0.501	0.001	0.952	0.012	0.012	0.012	0.012
GMM: rho weighted	0.988	-0.012	0.745	0.507	0.007	0.906	0.014	0.012	0.019	0.013
Mean z -score Index	0.987	-0.013	0.879	0.658	0.158	0.000	0.015	0.012	0.020	0.159
Factor Index 1	0.982	-0.018	0.000	0.567	0.067	0.000	0.014	0.012	0.023	0.068
IV ($z_2, z_3 \rightarrow z_1$)	0.953	-0.047	0.053	0.524	0.024	0.546	0.013	0.013	0.049	0.027
PLS Index	0.916	-0.084	0.000	0.567	0.067	0.000	0.013	0.012	0.085	0.068
Factor Index 2	0.871	-0.129	0.000	0.601	0.101	0.000	0.013	0.012	0.130	0.102
Factor Index 3	0.843	-0.157	0.000	0.635	0.135	0.000	0.029	0.020	0.160	0.136
Single Indicator z_1	0.652	-0.348	0.000	0.675	0.175	0.000	0.010	0.012	0.348	0.175
PCA Index	0.614	-0.386	0.000	0.601	0.101	0.000	0.009	0.012	0.387	0.102
IV ($z_1, z_2 \rightarrow z_3$)	1.383	0.383	0.000	0.656	0.156	0.005	0.062	0.034	0.388	0.160
Equal-weight Index	0.243	-0.757	0.000	0.535	0.035	0.157	0.003	0.012	0.757	0.037
Single Indicator z_2	0.099	-0.901	0.000	0.508	0.008	0.896	0.001	0.012	0.901	0.014
IV ($z_1, z_3 \rightarrow z_2$)	0.096	-0.904	0.000	0.520	0.020	0.617	0.001	0.012	0.904	0.024
Single Indicator z_3	0.089	-0.911	0.000	0.979	0.479	0.000	0.006	0.013	0.911	0.479
GMM	0.045	-0.955	0.000	0.507	0.007	0.907	0.015	0.012	0.956	0.013
All Indicators	0.045	-0.955	0.000	0.507	0.007	0.909	0.015	0.012	0.956	0.013

NOTES.—The table presents the performance of various estimation methods for β and γ in the model $y = \alpha + \beta x^* + \gamma w + \varepsilon$, where x^* is unobserved. Data were simulated with 10,000 observations and 1,000 repetitions. The true values are $\beta = 1$, $\gamma = 0.5$, and $\alpha = 0$. The table shows the average estimated coefficients, bias, coverage probability of 95% confidence intervals, standard deviation of the estimates, and root mean squared error (RMSE) for each parameter and method. The IV rows report 2SLS estimates using two indicators as instruments for the third, and the GMM rows report estimates using a generalized method of moments framework. Indices (PCA, PLS, equal-weight, and mean z -score) combine all indicators.



(A) β



(B) γ

FIGURE 2

SCENARIO 5: RMSE AS THE NUMBER OF INDICATORS VARY

NOTES.— Blah blah blah.

(Scenario 6.)

TABLE 5
SCENARIO 6: FORMATIVE INDICATORS MODEL

Method	$\hat{\beta}$	Bias (β)	Coverage (β)	$\hat{\gamma}$	Bias (γ)	Coverage (γ)	$\sigma_{\hat{\beta}}$	$\sigma_{\hat{\gamma}}$	RMSE (β)	RMSE (γ)
<i>Panel A. $\mathcal{J} = 5, J = 3$</i>										
True x^*	1.000	0.000	0.952	0.500	0.000	0.952	0.016	0.009	0.016	0.009
IV ($z_1, z_2 \rightarrow z_3$)	0.987	-0.013	0.418	0.360	-0.140	0.017	0.101	0.052	0.102	0.149
IV ($z_1, z_2 \rightarrow z_1$)	0.986	-0.014	0.404	0.361	-0.139	0.014	0.102	0.052	0.103	0.149
IV ($z_1, z_3 \rightarrow z_2$)	0.990	-0.010	0.424	0.359	-0.141	0.015	0.103	0.052	0.104	0.151
Mean z -score Index	0.838	-0.162	0.016	0.558	0.058	0.040	0.057	0.022	0.172	0.062
Factor Index 1	0.838	-0.162	0.000	0.558	0.055	0.050	0.057	0.022	0.172	0.062
Factor Index 3	0.789	-0.211	0.000	0.558	0.055	0.050	0.054	0.022	0.217	0.062
PLS Index	0.689	-0.311	0.000	0.556	0.056	0.050	0.048	0.022	0.315	0.060
Factor Index 2	0.684	-0.316	0.000	0.558	0.058	0.050	0.047	0.022	0.320	0.062
Equal-weight Index	0.592	-0.408	0.000	0.558	0.058	0.040	0.041	0.022	0.410	0.062
All Indicators	0.592	-0.408	0.000	0.558	0.058	0.037	0.041	0.022	0.410	0.062
GMM	0.592	-0.408	0.000	0.558	0.058	0.036	0.041	0.022	0.410	0.062
PCA Index	0.484	-0.516	0.000	0.558	0.058	0.040	0.033	0.022	0.517	0.062
Single Indicator z_1	0.330	-0.670	0.000	0.689	0.189	0.000	0.054	0.028	0.672	0.191
Single Indicator z_3	0.329	-0.671	0.000	0.689	0.189	0.000	0.055	0.028	0.673	0.191
Single Indicator z_2	0.327	-0.673	0.000	0.690	0.190	0.000	0.054	0.028	0.675	0.192
<i>Panel B. $\mathcal{J} = 20, J = 3$</i>										
True x^*	0.999	-0.001	0.955	0.500	0.000	0.964	0.019	0.009	0.019	0.009
IV ($z_1, z_3 \rightarrow z_2$)	0.778	-0.222	0.000	0.464	-0.036	0.388	0.037	0.020	0.225	0.041
IV ($z_1, z_2 \rightarrow z_3$)	0.777	-0.223	0.000	0.465	-0.035	0.407	0.036	0.020	0.226	0.041
IV ($z_1, z_2 \rightarrow z_1$)	0.777	-0.223	0.000	0.465	-0.035	0.400	0.038	0.021	0.227	0.041
Mean z -score Index	0.659	-0.341	0.000	0.620	0.120	0.000	0.024	0.011	0.342	0.121
Factor Index 1	0.659	-0.341	0.000	0.620	0.120	0.000	0.024	0.011	0.342	0.121
Factor Index 3	0.621	-0.379	0.000	0.620	0.120	0.000	0.023	0.011	0.379	0.121
PLS Index	0.539	-0.461	0.000	0.620	0.120	0.000	0.020	0.011	0.462	0.121
Factor Index 2	0.538	-0.462	0.000	0.620	0.120	0.000	0.020	0.011	0.462	0.121
All Indicators	0.466	-0.534	0.000	0.620	0.120	0.000	0.017	0.011	0.534	0.121
GMM	0.466	-0.534	0.000	0.620	0.120	0.000	0.017	0.011	0.534	0.121
Equal-weight Index	0.466	-0.534	0.000	0.620	0.120	0.000	0.017	0.011	0.534	0.121
PCA Index	0.381	-0.619	0.000	0.620	0.120	0.000	0.014	0.011	0.620	0.121
Single Indicator z_1	0.259	-0.741	0.000	0.724	0.224	0.000	0.016	0.011	0.741	0.224
Single Indicator z_3	0.259	-0.741	0.000	0.724	0.224	0.000	0.016	0.011	0.741	0.224
Single Indicator z_2	0.258	-0.742	0.000	0.724	0.224	0.000	0.016	0.011	0.742	0.224

NOTES.—The table presents the performance of various estimation methods for β and γ in the model $y = \alpha + \beta x^* + \gamma w + \varepsilon$, where x^* is unobserved. Data were simulated with 10,000 observations and 1,000 repetitions. The true values are $\beta = 1$, $\gamma = 0.5$, and $\alpha = 0$. The table shows the average estimated coefficients, bias, coverage probability of 95% confidence intervals, standard deviation of the estimates, and root mean squared error (RMSE) for each parameter and method. The IV rows report 2SLS estimates using two indicators as instruments for the third, and the GMM rows report estimates using a generalized method of moments framework. Indices (PCA, PLS, equal-weight, and mean z -score) combine all indicators. Given the data-generating process, the three IV estimators and the three single-indicator estimators are essentially identical; they should be viewed as a single estimator.

TABLE 6
SCENARIO XX: INDEX IS A NUISANCE FUNCTION ONLY

Reflective Indicators Models					Formative Indicators Models				
Method	$\hat{\gamma}$	Bias (γ)	$\sigma_{\hat{\gamma}}$	RMSE (γ)	Method	$\hat{\gamma}$	Bias (γ)	$\sigma_{\hat{\gamma}}$	RMSE (γ)
<i>Panel A. Correlated Indicators</i>					<i>Panel C. $\mathcal{J} = 3, J = 3$</i>				
True x^*	0.496	-0.004	0.023	0.024	True x^*	0.500	0.000	0.020	0.020
Matching + LASSO	0.669	0.169	0.026	0.171	Matching + LASSO	0.503	0.003	0.023	0.023
Matching	0.670	0.170	0.025	0.172	LASSO	0.510	0.010	0.021	0.023
LASSO	0.677	0.177	0.024	0.179	Matching	0.504	0.004	0.023	0.023
<i>Panel B. Correlated Indicators, One Indicator is Perfect</i>					<i>Panel D. $\mathcal{J} = 5, J = 3$</i>				
True x^*	0.503	0.003	0.025	0.026	True x^*	0.502	0.002	0.020	0.020
Matching + LASSO	0.509	0.009	0.026	0.028	Matching + LASSO	0.536	0.036	0.022	0.042
Matching	0.513	0.013	0.026	0.029	Matching	0.536	0.036	0.022	0.042
LASSO	0.519	0.019	0.027	0.032	LASSO	0.541	0.041	0.021	0.046
					<i>Panel E. $\mathcal{J} = 20, J = 3$</i>				
					True x^*	0.502	0.002	0.018	0.018
					Matching + LASSO	0.565	0.065	0.020	0.068
					Matching	0.566	0.066	0.020	0.069
					LASSO	0.576	0.076	0.018	0.078

NOTES.—

7 Replication: Ortoleva and Snowberg (2015)

Alfredo write-up? :)

TABLE 7
REPLICATION: SENSITIVITY TO CREATION OF MEDIA INDEX

	PCA	Mean z -Score	PLS	LW	PCA	Mean z -Score	PLS	LW
<i>Panel A. Overconfidence</i>								
Media Index	0.205*** (0.041)	0.338*** (0.065)	0.210*** (0.039)		0.149*** (0.038)	0.236*** (0.060)	0.181*** (0.037)	
Blog				0.025 (0.066)				0.053 (0.071)
TV				0.271*** (0.085)				0.127* (0.073)
Newspaper				0.169*** (0.055)				0.073 (0.054)
Radio				0.254*** (0.067)				0.246*** (0.062)
LW Estimate				0.719*** (0.143)				0.500*** (0.132)
LW SE								
Controls	N	N	N	N	Y	Y	Y	Y
Observations	2927	2927	2927	2927	2927	2927	2927	2927
<i>Panel B. Ideology</i>								
Media Index	0.059** (0.023)	0.095** (0.039)	0.187*** (0.031)		0.055** (0.022)	0.079** (0.036)	0.189*** (0.026)	
Blog				-0.139*** (0.049)				-0.087* (0.051)
TV				0.071 (0.066)				-0.001 (0.061)
Newspaper				-0.141*** (0.047)				-0.167*** (0.046)
Radio				0.374*** (0.056)				0.385*** (0.052)
LW Estimate				0.165* (0.086)				0.130* (0.075)
LW SE								
Controls	N	N	N	N	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868
<i>Panel C. Squared Deviation</i>								
Media Index	0.288*** (0.028)	0.430*** (0.044)	0.294*** (0.028)		0.189*** (0.028)	0.281*** (0.044)	0.209*** (0.024)	
Blog				0.346*** (0.069)				0.270*** (0.056)
TV				-0.002 (0.059)				-0.008 (0.059)
Newspaper				0.263*** (0.060)				0.147*** (0.051)
Radio				0.313*** (0.056)				0.200*** (0.044)
LW Estimate				0.920*** (0.100)				0.609*** (0.101)
LW SE								
Controls	N	N	N	N	Y	Y	Y	Y
Observations	2868	2868	2868	2868	2868	2868	2868	2868

NOTES.— PCA = Principal Component Analysis. LW = Lubotsky and Wittenberg (2006). N = Number of observations. PCA columns are identical to Table 2 in Ortoleva and Snowberg (2015). Standard errors clustered by age. * $p < .10$, ** $p < .05$, *** $p < .01$.

TABLE 8
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
<i>Panel A. Ideology</i>												
Overconfidence Index	0.218*** (0.027)	0.237*** (0.030)	0.198*** (0.024)	-0.011 (0.025)	0.220*** (0.023)	0.240*** (0.026)	0.200*** (0.021)	-0.008 (0.025)	0.199*** (0.023)	0.218*** (0.026)	0.181*** (0.021)	0.000 (0.025)
Reported Unemp				0.098*** (0.036)				0.110*** (0.034)				0.110*** (0.033)
Reported Inflation				0.017 (0.037)				0.001 (0.033)				-0.001 (0.032)
Expected Unemp				0.110*** (0.037)				0.112*** (0.036)				0.087*** (0.034)
Expected Inflation												
LW Estimate				0.215*** (0.029)				0.215*** (0.025)				0.196*** (0.025)
LW SE												
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
<i>Panel B. Ideological Extremeness Purged of Economic Controls</i>												
Overconfidence Index	0.234*** (0.028)	0.259*** (0.031)	0.208*** (0.025)	0.076** (0.034)	0.174*** (0.027)	0.192*** (0.030)	0.155*** (0.024)	0.038 (0.033)	0.122*** (0.026)	0.135*** (0.029)	0.109*** (0.023)	0.029 (0.029)
Reported Unemp				0.063 (0.038)				0.065* (0.035)				0.046 (0.031)
Reported Inflation				0.068 (0.043)				0.025 (0.038)				0.028 (0.035)
Expected Unemp				0.042 (0.042)				0.053 (0.038)				0.024 (0.034)
Expected Inflation												
LW Estimate				0.249*** (0.030)				0.181*** (0.029)				0.128*** (0.029)
LW SE												
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. LW = Lubotsky and Wittenberg (2006). N = Number of observations. PCA columns are identical to Table 3 in Ortoreva and Snowberg (2015). Standard errors clustered by age. * p < .10, ** p < .05, *** p < .01.

8 Application

American (US) politics has entered an era of pronounced geographic polarization (McCarty et al. 2019). While the partisan divide between urban and rural areas has received substantial attention, far less is known about how population health correlates with recent shifts in presidential voting at the level of individual congressional districts (CDs). Wasfy et al. (2020) show that, across US counties, a composite index of adverse health conditions predicts changes in Republican vote share between the 2016 and 2018 elections. Their work raises an important question for economists interested in the political-economy consequences of health inequality: does worse health systematically tilt electoral sentiment toward one or the other (Republican or Democrat) presidential candidate?

In this application of the methods discussed above, we answer that question by focusing on the vote-share shift for the Republican presidential nominee between 2020 and 2024. Unlike counties, CDs are the fundamental units of representation in the US House of Representatives and are therefore the natural jurisdiction for congressional policymaking. Using data from the Congressional District Health Dashboard (CDHD) and precinct-aggregated presidential returns from The Downballot project, we assemble a district-level panel that links more than two dozen health metrics to two-party vote shares in 2020 and 2024.

We distill the rich CDHD information into a single health index using dimensionality-reduction techniques described in Sections 3.1, 4.1, and 4.3 as well as the approach that does not require index creation. Examples of health indicators in the dataset include deaths from a variety of causes; prevalence rates of diabetes, low birthweight, obesity, and so on; as well as measures of air pollution, housing with potential lead risk, and food insecurity.

The main outcome variable is the district-level change in Republican two-party vote share: $\Delta \text{Vote}_i = \text{RepShare}_{i,2024} - \text{RepShare}_{i,2020}$. To quantify the association between health and partisan realignment, we estimate the following model:

$$\Delta \text{Vote}_i = \alpha + \beta_i H_i + \gamma' \mathbf{X}_i + \delta_s + \varepsilon_i, \quad (35)$$

where \mathbf{X}_i is a vector of socioeconomic and demographic controls measured prior to the 2020 election (e.g., median household income, educational attainment, population density, and racial or ethnic composition), δ_s is a vector of state fixed effects, and ε_i is a disturbance term. Standard errors are clustered at the state level to account for within-state correlation in both electoral and health environments.

Election data from MIT Election Lab (Data and Lab 2018) Ideology data from (Tausanovitch and Warshaw 2013; Warshaw and Tausanovitch 2022) Health data from (Wisconsin Population Health Institute 2025)

9 Conclusion

Future work to examine case of the index as the dependent variable.

References

- Akee, Randall et al. (2018). “How Does Household Income Affect Child Personality Traits and Behaviors?” *American Economic Review* 108.3, pp. 775–827. DOI: [10.1257/aer.20160133](https://doi.org/10.1257/aer.20160133).
- Andersson, Jonas and Jarle Møen (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. DOI: <https://doi.org/10.1111/obes.12103>.
- Ash, Elliott, Sergio Galletta, and Tommaso Giommoni (2025). “A Machine Learning Approach to Analyze and Support Anticorruption Policy”. *American Economic Journal: Economic Policy* 17, pp. 162–93. DOI: [10.1257/pol.20210618](https://doi.org/10.1257/pol.20210618).
- 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. DOI: <https://doi.org/10.1016/j.econlet.2015.09.029>.
- Black, Dan A, Mark C Berger, and Frank A Scott (2000). “Bounding Parameter Estimates with Non-classical Measurement Error”. *Journal of the American Statistical Association* 95.451, pp. 739–748. DOI: [10.1080/01621459.2000.10474262](https://doi.org/10.1080/01621459.2000.10474262).
- 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. DOI: [10.1093/qje/qjt057](https://doi.org/10.1093/qje/qjt057).
- 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. DOI: [10.1177/1094428106294734](https://doi.org/10.1177/1094428106294734).
- Bollen, Kenneth A (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Incorporated.
- 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. DOI: [10.1162/003465303322369722](https://doi.org/10.1162/003465303322369722).
- 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. DOI: [10.1016/j.jeconom.2015.02.013](https://doi.org/10.1016/j.jeconom.2015.02.013).
- 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. DOI: [10.1162/REST_a_00139](https://doi.org/10.1162/REST_a_00139).

- Data, MIT Election and Science Lab (2018). *County Presidential Election Returns 2000-2020*. Version V13. DOI: [10.7910/DVN/VOQCHQ](https://doi.org/10.7910/DVN/VOQCHQ).
- 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. DOI: [10.1007/978-3-540-32827-8_2](https://doi.org/10.1007/978-3-540-32827-8_2).
- 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. DOI: [10.1093/jrsssa/qnae069](https://doi.org/10.1093/jrsssa/qnae069).
- Filmer, Deon and Kinnon Scott (2012). “Assessing Asset Indices”. *Demography* 49, pp. 359–392. DOI: [10.1007/s13524-011-0077-5](https://doi.org/10.1007/s13524-011-0077-5).
- 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. DOI: [10.1086/701681](https://doi.org/10.1086/701681).
- Griliches, Zvi (1977). “Estimating the Returns to Schooling: Some Econometric Problems”. *Econometrica* 45.1, pp. 1–22.
- (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. DOI: <https://doi.org/10.1016/B978-0-08-091857-0.50015-7>.
- Hyslop, R and Guido W Imbens (2001). “Bias from Classical and Other Forms of Measurement Error”. *Journal of Business & Economic Statistics* 19.4, pp. 475–481.
- 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. DOI: [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202).
- Kim, Dongwoo and Daniel Wilhelm (2024). “Powerful t-tests in the presence of nonclassical measurement error”. *Econometric Reviews* 43.6, pp. 345–378. DOI: [10.1080/07474938.2024.2334166](https://doi.org/10.1080/07474938.2024.2334166).
- Kippersluis, Hans van and Cornelius A Rietveld (2018). “Beyond plausibly exogenous”. *The Econometrics Journal* 21.3, pp. 316–331. DOI: [10.1111/ectj.12113](https://doi.org/10.1111/ectj.12113).

- Kiviet, Jan F (2016). “When is it really justifiable to ignore explanatory variable endogeneity in a regression model?” *Economics Letters* 145, pp. 192–195. doi: <https://doi.org/10.1016/j.econlet.2016.06.021>.
- (2020). “Testing the impossible: Identifying exclusion restrictions”. *Journal of Econometrics* 218.2, pp. 294–316. doi: <https://doi.org/10.1016/j.jeconom.2020.04.018>.
- (2023). “Instrument-free inference under confined regressor endogeneity and mild regularity”. *Econometrics and Statistics* 25, pp. 1–22. doi: <https://doi.org/10.1016/j.ecosta.2021.12.008>.
- 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. doi: <https://doi.org/10.1016/j.jeconom.2009.05.005>.
- Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007). “Experimental Analysis of Neighborhood Effects”. *Econometrica* 75.1, pp. 83–119.
- 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. doi: [10.1177/1536867X211045575](https://doi.org/10.1177/1536867X211045575).
- Lewbel, Arthur (2012). “Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models”. *Journal of Business & Economic Statistics* 30.1, pp. 67–80. doi: [10.1080/07350015.2012.643126](https://doi.org/10.1080/07350015.2012.643126).
- 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. doi: [10.1080/07350015.2023.2166052](https://doi.org/10.1080/07350015.2023.2166052).
- Lubotsky, Darren and Martin Wittenberg (2006). “Interpretation of Regressions with Multiple Proxies”. *Review of Economics and Statistics* 88.3, pp. 549–562. doi: [10.1162/rest.88.3.549](https://doi.org/10.1162/rest.88.3.549).
- 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. doi: [10.1257/aer.99.3.1006](https://doi.org/10.1257/aer.99.3.1006).
- McCarty, Nolan et al. (2019). “Geography, uncertainty, and polarization”. *Political Science Research and Methods* 7.4, pp. 775–794.
- 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. doi: <https://doi.org/10.1093/ajae/aay029>.
- 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. doi: [10.1257/aer.20211094](https://doi.org/10.1257/aer.20211094).

- 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. DOI: [10.1162/REST_a_00171](https://doi.org/10.1162/REST_a_00171).
- Ortoleva, Pietro and Erik Snowberg (2015). “Overconfidence in Political Behavior”. *The American Economic Review* 105.2, pp. 504–535.
- 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. DOI: <https://doi.org/10.1111/roiw.70016>.
- Ravallion, Martin (2012). “Mashup Indices of Development”. *The World Bank Research Observer* 27.1, pp. 1–32. DOI: [10.1093/wbro/lkr009](https://doi.org/10.1093/wbro/lkr009).
- 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. DOI: <https://doi.org/10.1016/j.paid.2015.07.019>.
- Samuelson, Paul A (1983). *Foundations of Economic Analysis*. Harvard University Press.
- Solon, Gary (1992). “Intergenerational Income Mobility in the United States”. *The American Economic Review* 82.3, pp. 393–408.
- Stoetzer, Lukas F, Xiang Zhou, and Marco Steenbergen (2025). “Causal inference with latent outcomes”. *American Journal of Political Science* 69.2, pp. 624–640. DOI: <https://doi.org/10.1111/ajps.12871>.
- Tausanovitch, Chris and Christopher Warshaw (2013). “Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities”. *The Journal of Politics* 75.2, pp. 330–342. DOI: [10.1017/S0022381613000042](https://doi.org/10.1017/S0022381613000042).
- Warshaw, Christopher and Chris Tausanovitch (2022). *Subnational ideology and presidential vote estimates (v2022)*. Version V1. DOI: [10.7910/DVN/BQKU4M](https://doi.org/10.7910/DVN/BQKU4M).
- 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.
- Wisconsin Population Health Institute, University of (2025). *County Health Rankings & Roadmaps*. URL: www.countyhealthrankings.org.
- Wold, H O A (1982). “Soft modelling: the basic design and some extensions”. In: *Systems under indirect observation, Part II*. Jöreskog, K G and Wold, H O A (eds.) North Holland.

- 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. DOI: [10.1111/obes.12536](https://doi.org/10.1111/obes.12536).
- Young, Alwyn (2022). “Consistency without Inference: Instrumental Variables in Practical Application”. *European Economic Review* 147, p. 104112. DOI: <https://doi.org/10.1016/j.euroecorev.2022.104112>.
- Zhang, Jeffrey and Junu Lee (2025). “A general condition for bias attenuation by a nondifferentially mismeasured confounder”. *Biometrika*, asaf026. DOI: [10.1093/biomet/asaf026](https://doi.org/10.1093/biomet/asaf026).