Series Name: WPS
Paper No.: 261
Issue Date: May 1, 2026

Measuring What Matters: A Guide to Index Creation in Applied Microeconomics

Bruce Wydick

Recommended Citation:

Wydick, Bruce (2026): Measuring What Matters: A Guide to Index Creation in Applied Microeconomics. CEGA Working Paper Series No. WPS-261. Center for Effective Global Action. University of California, Berkeley. Text. https://doi.org/10.26085/C3N01J.

Measuring What Matters: A Guide to Index Creation in Applied Microeconomics

Bruce Wydick[1,2,3,4]

May 1, 2026

Keywords: Index Construction, Welfare Measures, Development Indices, Data Quality

Abstract

Research in development, labor, and health economics increasingly relies on indices to summarize the complex dimensions of human well-being—such as health, wealth, aspirations, and social inclusion—into single measures. Yet too often index construction is often driven by convention and precedent rather than by a clear conceptual framework, leaving little guidance on selecting the appropriate type of index for a given context, the data used to construct it, or the variables included. This paper provides a guide to index construction in applied microeconomics. It categorizes different index types and their appropriate uses, clarifies the tradeoffs between measurement and statistical power, and offers a decision tree to guide selection among approaches such as factor analysis, principal components, and aggregation-based indices. The paper also outlines principles for variable selection via machine learning in high-dimensional settings and concludes with an example of index construction in a three-country randomized trial with an entrepreneurial aspirations intervention. By linking conceptual foundations to applied practice, the paper aims to improve the transparency and rigor of index-based measurement in applied research.

Affiliations: [1]Professor of Economics, University of San Francisco, [2]Adjunct Professor, University of California at Davis; [3]Distinguished Research Affiliate, University of Notre Dame, Kellogg Institute for International Studies and CEIDS; [4]Research Affiliate, Center for Effective Global Action, University of California at Berkeley; Email: wydick@usfca.edu, Department of Economics, University of San Francisco, San Francisco, CA 94117.

1 Introduction

Economics has expanded its ambitions, increasingly measuring not just what people possess, but broader constructs of human wellbeing. We now ask not only whether people are richer, but whether they feel capable, hopeful, resilient, socially included, and free. Where traditional welfare indicators focused on income, education, and health, economics has turned increasingly toward more abstract outcomes — psychological constructs, socio-economic status, and broader aggregations of human flourishing (Banerjee and Duflo, 2013; Stiglitz, 2020; Jansen et al., 2024).

This is true particularly in development economics, from which the present paper will draw many of its examples. But it is also true in other applied fields: labor economics in measuring skill requirements and task complexity (Autor et al., 2013; Deming and Kahn, 2021); behavioral economics in measuring different types of preferences, traits, and values (Haushofer and Shapiro, 2016; Chinoy et al., 2024); health economics in quality-of-life indices (Brazier and Rowen, 2021; Zrubka et al., 2022); organizational behavior in assessing organizational and work characteristics (Edmondson, 1999; Bloom et al., 2020); agricultural economics in assessing factors such as crop suitability and substitutability (Olsson and Hibbs, 2005; Galor and Ozak, 2016); and environmental economics in summary valuations for environmental goods (Ziegler, 2021; Vásquez-Lavín et al., 2021). Based on searches in the Econ Lit database, Figure 1 shows the increase in journal articles across the years 2001 to 2025 in the top five general economics journals and the top five development economics journals whose paper titles included eight leading abstract constructs.[1]

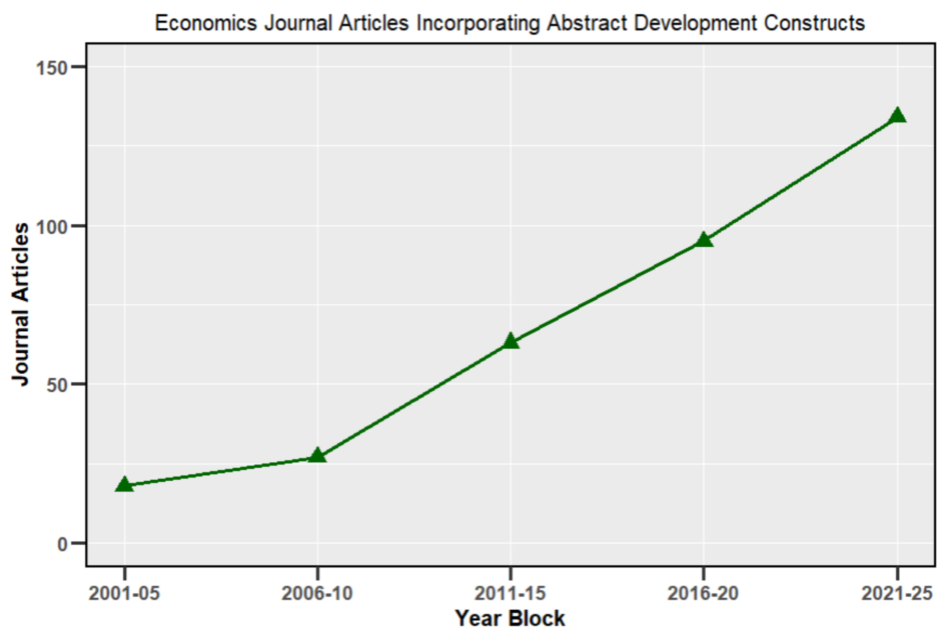
Because these aggregated and abstract outcomes are not directly quantifiable in the same manner as traditional economic outcomes, researchers typically measure them through the creation of *indices*, which aggregate information across multiple indicators. Yet even as the field's ambitions have grown, its scrutiny of the indices used to reflect these abstract constructs has lagged behind. Applied researchers agonize over causal identification, yet routinely leave choices over index construction unjustified — even though, as this paper will show, a poorly constructed index can bias treatment effect estimates on latent constructs just as severely. Which type of index to use and why in various contexts is not always clear and can be a vexing exercise. This article seeks to provide a guide for the construction of indices used to measure these important but

[1] These ten journals are the *American Economic Review, Journal of Political Economy, Quarterly Journal of Economics, Econometrica, Review of Economic Studies, Journal of Development Economics, World Development, World Bank Economic Review, Economic Development and Cultural Change, and the Journal of Development Studies.* Search was for paper titles that contained one of the following words: *subjective well-being, aspirations, empowerment, resilience, agency, perseverance, self-control, social inclusion*.

more abstract dimensions of human experience.

Indices are used as nearly every kind of variable in empirical research. Psychological and sociological constructs are frequently used as dependent variables to estimate broader causal effects of interventions. For example, randomized evaluations of cash transfers often document improvements in psychological well-being and reductions in stress and depression among poor households (Haushofer and Shapiro, 2016; Ridley et al., 2020a). Studies of conflict and violence likewise study impacts on trauma, post-traumatic stress disorder (PTSD), and resilience as key dimensions of welfare affected by war and displacement (Blattman and Annan, 2010; Voors et al., 2012; Baird et al., 2022). Researchers use *summary indices* as dependent variables to look at overall effects, address problems of over-testing, and to increase statistical power (Anderson, 2008; Kling et al., 2007).

Figure 1: Increase in Abstract Constructs in Titles of Journal Articles



Indices are also used as independent variables, such as those that control for wealth or socio-economic status (SES) (Filmer and Pritchett, 2001a; Chetty et al., 2016). As psychological constructs, they also appear in economic research as mechanisms shaping economic behavior. A growing literature emphasizes the role of aspirations, empowerment, and agency in influencing forward-looking decisions such as in education and microenterprise development (Dalton et al., 2016; Genicot and Ray, 2017; Lybbert and Wydick, 2018a). Empirical studies in rural Ethiopia, for example, show that variation in aspirations can influence households' economic choices and investment behavior (Bernard et al., 2014). These developments reflect a broader recognition that poverty interacts with psychological processes—including hope, aspirations, and mental health—that shape economic decision-making and long-run development trajectories (Lybbert and Wydick, 2018a).

Still other research has used psychological constructs as treatment variables. Here interventions seek to elevate a given psychological construct that in turn is hoped to directly influence behavior auspicious to elevated development outcomes such as better school performance, physical health, or microenterprise growth. Programs incorporating cognitive behavioral therapy, mentorship, or mindset training aim to improve outcomes such as self-control, resilience, and expectations about the future (Blattman et al., 2017; Rojas Valdes et al., 2022; Riley, 2024; Bernard et al., 2026). However, even as indices are employed for myriad uses in empirical research, their use is too often based on convention and precedent, often lacking a clear rationale and framework.

In Section 2 of this paper I will clarify important terms related to the classification of indices. In Section 3 I review different types of commonly used indices and present a decision tree to help guide index choice. In Section 4 I provide a rationale for different data types that are most appropriate in different types of indices. A further question exists over variable selection within indices, and in Section 5 I present LASSO and elastic net-based algorithms to be used for pruning formative and reflective indices, respectively. In Section 6 I provide an application of these principles of index creation to a separate working paper that contains results of a three-country pilot study in which we introduced entrepreneurial aspirations and financial literacy interventions in a cross-cut randomized trial in India, Peru, and Uganda. Section 7 concludes with general recommendations for index creation in applied research.

2 Terminology and Fundamental Index Distinctions

A first step to systematic index construction is understanding the different classes of indices. A primary distinction is between indices designed to measure latent constructs and indices intended to aggregate multiple dimensions of welfare.

2.1 Construct Indices vs. Welfare Indices

A primary source of confusion in index creation is in the distinction between *construct* and *welfare* indices. This misunderstanding is likely rooted in the inter-disciplinary communication breakdowns that naturally occur within the many research areas that are simultaneously inhabited by economists, public health researchers, sociologists, anthropologists, political scientists, and psychologists.

The distinction between these two broad types of indices originates from the different empirical traditions that developed within psychology and economics. Because most phenomena studied in psychology, such as resilience, agency, and depression—are not directly observable quantities, empirical psychology evolved largely as a science built around measuring latent "constructs" (Cronbach and Meehl, 1955; Borsboom, 2005; Furr, 2021). As a result, psychometrics has invested more thought and time than economics in the creation and measurement of abstract constructs. This is typically done through the assimilation of multiple responses to a questionnaire or multiple behavioral observations. Tools to create and measure psychological constructs such as principal components, and especially factor analysis, have consequently become central tools of the discipline.

Economics, in contrast, did not evolve as strongly as a measurement science because it was traditionally concerned with more easily measurable phenomena: inputs, outputs, prices, profits, and other market outcomes. As a result, the discipline historically devoted less attention to measurement and more attention to aggregating these observable outcomes into welfare-relevant summaries. Indices in economics created to measure economic welfare thus typically combine variables that are themselves interpretable as components of material well-being, such as income, consumption, education, or health. The purpose of economic welfare indices is not to measure a latent construct but rather to summarize welfare across multiple outcome dimensions.

2.2 Reflective vs. Formative Indices

A second distinction in index construction is between *reflective* and *formative* indices. This distinction is especially important with respect to the nature of the causal relationship between the index and its component variables.

A *reflective* index is one in which causality runs from a latent construct to a set of observable data that *reflect* the latent construct. The causal relationship for reflective indices is *latent construct* $\rightarrow$ *observable indicators*. The construct thus "reflects" (causes) observable behavior. A development economist, for example, might want to recover the observable indicators of a latent "growth mindset" construct. One characteristic of growth-mindset individuals is that they typically seek and enjoy challenges rather than avoid them. Thus a growth mindset yields a set of observable indicators related to persisting in solving a problem instead of quickly giving up. One item in the index thus might include the number of attempts and/or strategies used to solve an experimental puzzle. A second characteristic of growth mindset is positive self-talk. So another item in the index might include the extent to which during the puzzle exercise, a subject articulates statements such as "I know I can get this."

The point of a reflective index is that the observable outcome, whether observed behavior or response in a questionnaire, is that the k^{th} variable in the index for individual i, x_{ki}, is a function of the underlying construct and an error, or $x_{ki} = \lambda_k \theta_i + \epsilon_{ki}$. The latent construct θ_i in a reflective index is "creating" the x_{ki} that are observable to the researcher.

The causality runs in the opposite direction with a formative index: *observable indicators* $\rightarrow$ *formative index*. An example here is the creation of a SES (socio-economic status) index, often comprised of variables reflecting income, education, and leadership, or $SES = \Phi(I, E, L)$. Here the outcomes "create" the index, rather than latent construct "creating" the outcomes. Poverty measurement, dwelling quality indices, and the human development indicators are other examples of formative indices. Unlike reflective indices, which emanate from a single and stable latent characteristic, if we remove one component from formative index, it changes the nature of the index.

Most welfare indices aggregate outcomes that are *formative* rather than *reflective*, such as the Human Development Index (HDI), where the HDI is an equally weighted index of measures of education, health, and income. Most multi-dimensional poverty indices treat such indicators not as reflections of an underlying latent variable, but as dimensions that collectively constitute the construct itself. Likewise most construct indices are reflective, such as an indices that capture latent anxiety or depression.

Important exceptions exist. An index of financial literacy is a construct index, but it is made up of formative components like financial numeracy, confidence in financial decisions, and understanding of interest rates. Here a *construct* index is *formative*; our choice of index components alters the definition and meaning of the index. We can also view some *welfare* indices as *reflective*, such as food insecurity. Here a latent welfare construct (food insecurity) *reflects* (causes) worry about food, going hungry on some days, and missed meals.

2.3 Statistically vs. Conceptually Weighted Indices

Critical to any index is the determination of weights for the items it includes. A third distinction lies between *statistically* vs. *conceptually* weighted indices, the latter of which stem from theory or normative judgment. For example, the HDI index assigns equal importance to health, education, and income by implicitly assigning each a weight of one-third.

The decision about whether to hand over the steering wheel to the data in assigning index weights, or whether to impose them *a priori*, on the data is often one of the most important choices a researcher makes in index creation. But too often this research choice is made arbitrarily or by tradition, handed down like a family recipe, when there actually exist clear guidelines in selecting a choice.

Statistical (data-driven) approaches to component weight assignment assume a reflective model in which a latent construct drives the measurable outcomes observed by researchers. Indices that measure agency provide a useful illustration of how these different approaches arise in behavioral economics research. Traditionally the agency of an individual is regarded as an underlying latent construct reflecting individuals' ability to make and act upon choices that affect their lives. Such an index might be created using questions related to perceived autonomy, confidence in decision-making, and expectations about personal influence over future outcomes. These may then be interpreted as reflective indicators of the latent agency construct. In this formulation, psychometric methods will be used to create the index with variable weights based on their correlation with an index comprised of mutually correlated variables (Borsboom, 2005; Vyas and Kumaranayake, 2006). These methods implicitly assume a reflective measurement structure in which the underlying construct generates the observed indicators.

Much work in development, labor, and behavioral economics has either explicitly or implicitly assumed that agency is a multidimensional concept composed of several distinct capabilities. These could include specific indicators for control over one's environment, resources, and future, freedom of mobility, participation in household decisions, and the ability to pursue desired economic activities. Such outcomes may then be aggregated into an agency index using conceptually chosen weights. In such a case the indicators would be *formative* rather than *reflective*. In such a case they would purposefully and collectively define the concept of agency (perhaps to a specific context) rather than assume that they arise from a single underlying latent construct. They might even be assigned equal weights, reflecting their equal importance (or perhaps the researcher's willingness to explicitly value one aspect of agency above others). Measures of agency frequently have adopted this perspective (Kabeer, 1999; Alkire et al., 2013a; Bandiera et al., 2020; Rojas Valdes et al., 2022); but generally we want to let latent factors drive reflective index weights.

In *formative* indices correlation between index components is not important. Moreover, data-driven weighting methods are often inappropriate; the covariance structure of the indicators has no bearing on their importance. Instead, weights reflect *a priori* normative judgments about their relative contribution in a summary index. Proper index construction is conscious of the use of data-driven or theoretical weights and to thoughtfully choose one approach or the other where appropriate in context.

3 Common Index Types Used in Development Research

A handful of approaches dominate index creation in applied microeconomics, each embodying different assumptions about how component variables relate to the underlying construct of interest. In deciding which index is appropriate for a given context, it is critical to understand distinctions between construct vs welfare indices, reflective vs formative, and data-driven vs. conceptually driven weighting schemes. But other considerations matter, such as the primal importance of measurement or statistical power, the existence (and importance) of heterogeneous correlation between index variables, and the degree of substitutability between them.

3.1 The Straight Scale Index

The simplest approach to composite variable aggregation is a straight scale index. Used most commonly in psychology, a straight scale index orients each component variable so that higher values reflect better outcomes (or at least a unidirectional outcome), then sums the raw scores. A researcher might have five Likert-scale items measuring life satisfaction–reported depression, anxiety, aspirations, happiness, and optimism. A straight scale index first reverses the negatively-valenced items in line with the *name* of the index. (Outcomes in a "Depression Index" should not be higher for *less*-depressed people.) Then it adds the resulting scores for each individual x_{ik} such that

$$\text{Straight Scale Index}_i = \textstyle\sum_{k=1}^{K} x_{ik}$$

The procedure is transparent and intuitive. Its weakness is equally transparent and intuitive: even when questions contain equal numerical ranges, items with greater variance will dominate the index simply by virtue of their higher variation, not because they carry more substantive information about the underlying construct. This leads to an interesting question: Why would anybody use a straight-scale index?

The answer is that when variables are measured on comparable scales, when variances within items are approximately equal, and when the researcher has no reason to doubt their comparability, a straight scale index is defensible. The PHQ-9 (Patient Health Questionnaire-9), for example, is a globally used straight-scale depression screening instrument major depressive disorder used both in clinical care and research. Similarly, the GAD-7 (Generalized Anxiety Disorder-7) is one of the most widely used instruments for diagnosing generalized anxiety, consisting of seven straight-scale items mapping onto core symptoms.

More generally, straight-scale indices are common in psychology for a number of reasons. First, scaled indices, to some extent at least, were designed with questions intended to have approximately equal variance. And where some items in an index may be relatively rare but important, such as suicidal thoughts, they represent important outcomes whose importance could be distorted through item-scaling. The goal of screeners, therefore, is clinical significance, not statistical equalization. However, there does appear to exist some path dependence in their use, where coordination to a straight scale has occurred on an inferior equilibrium like the QWERTY keyboard. Raw scale totals represent clinically interpretable severity levels that have historical precedent, allowing for comparability with historical studies. Researchers use these scales because previous research used them. And they are utilized in conjunction with comfortably familiar numeric cutoffs used by clinicians, e.g., GAD-7 $\geq$ 10 implies a high likelihood of generalized anxiety disorder.

Applied economists should be aware of these indices, and given their level of validation, should use them when appropriate. But they should simultaneously be aware that in most applications, straight scale indices are unlikely to be optimal for capturing abstract phenomena outside of their more narrow applications in psychology.

3.2 The Kling Index

The Kling et al. index (Kling et al., 2007) addresses the general deficiencies of the straight scale through standardization, and over two decades it has become the workhorse index of applied microeconomic indices. Each component variable of a Kling index x_{ik} is standardized before aggregation, ensuring that no single variable dominates the index merely because of its numerical range. The standardized variables $z_{ik} = Z(x_{ik})$ are then summed and the resulting index is itself re-standardized, placing it on a common scale that facilitates comparison across outcomes and studies as given by

$$\text{Kling Index}_i = Z(\textstyle\sum_{k=1}^{K} z_{ik})$$

where $Z(\cdot)$ is the standardization function. This final re-standardization is important on practical terms: effect sizes on development outcomes are routinely interpreted against informal benchmarks, with impacts below 0.20 standard deviations in areas such as education being regarded as modest at best, those between 0.20 and 0.40 as substantively meaningful, and those above 0.40 typically regarded as large, though these thresholds certainly vary with context and outcome domain.

The Kling index is most appropriate when the researcher has tangible, directly-observed component variables for which there is no compelling theoretical reason to assign differential weights — dimensions of housing quality, for instance, or a set of physical health indicators each representing an independent facet of the same broader outcome. Its deliberate construction and implicit assumption of equal weighting puts it strongly in the *formative* and *conceptually weighted* index camp. This is a limitation but also a virtue; it forces the researcher to be explicit about the normative judgment that all components matter equally.

3.3 The Anderson Index

The Anderson index (Anderson, 2008) retains the standardization logic of the Kling approach but introduces a refinement that is consequential when component variables exhibit meaningful and heterogeneous correlation with one another. When multiple variables carry overlapping information about the same underlying dimension, equal weighting effectively double-counts that dimension. The Anderson index corrects for this by weighting each variable inversely by its covariance with the other components of the index: variables that contribute unique information receive higher weight, while those that are redundant with other measures receive lower weights. This weighting feature increases statistical power for dependent-variable measures of impact.

As a Generalized Least Squares (GLS) index, the Anderson weights are derived from the inverse of the covariance matrix Σ of the standardized component variables z_k. After standardizing the sum of the product of the component variables and their Anderson weights, we obtain

$$\text{Anderson Index}_i = Z((\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}'\Sigma^{-1}z_{ik}).$$

The Anderson Index is particularly well-suited to measuring impacts on aggregated outcomes such as an index of microenterprise performance, where revenues and profits may be more strongly correlated components of enterprise health, but other variables such as leverage ratios and formalization (obtaining the necessary legal permits for operation) might display more independent variation.

The Anderson index is appropriate when creating summary indices of performance in impact studies where there is "meaningful heterogeneous redundancy" within the variables in a formative index such that it becomes important to magnify the influence of individual index components that represent important and unique facets of welfare. Accounting for this redundancy improves statistical power, and so the Anderson index is useful as a dependent variable in impact studies as a dependent variable that summarizes treatment effects across multiple outcomes (e.g., Schwab et al., 2020). But its data-driven nature makes it inappropriate for weighting outcomes in a normative welfare index. For example, health and education may be considered to be two conceptually important components of an index; merely because they are highly correlated does not mean they should be "weighted down".

Interestingly, the Anderson Index is a *formative* index by nature and intention, and while the purpose of an Anderson index is not to recover latent psychological constructs, it is sometimes used in impact evaluation to identify effects on an aggregated set of variables intended to measure latent psychological outcomes when statistical power is of primal importance. But when doing so, it is not identifying an impact on the psychological construct itself, as would factor analysis, but rather a (potentially) high-powered index that is related to that construct.

3.4 Principal Components Index

Principal Component Analysis (PCA) takes a fundamentally different approach to the index construction problem. Where the Kling index begins with a normative judgment about which variables to include and that they should be weighted equally, PCA asks the data itself to reveal the underlying weighting structure. And where the Anderson index places higher weights on variables that display lower correlation with other variables in the index, PCA assigns *higher* weights to these variables.

PCA generates a set of "components" up to the dimensions of the index. The first principal component identifies the linear combination of variables that accounts for the maximum variance in the dataset. Geometrically, it traces the direction through variable space for which the K-dimensional cloud of observations is most widely dispersed. The second component then captures the largest remaining source of variation subject to the constraint that it is orthogonal to the first, and so on. Each successive component is uncorrelated with those that precede it, so the components together partition the total variance in the data into non-overlapping pieces. The PCA index is

$$PC_i = \sum_{k=1}^{K} w_k^p z_{ik}$$

where PC_i is the principal component index for observation i, z_{ik} is the standardized value of variable k for observation i, and w_k^p is the weight assigned to variable k by the eigenvector associated with the largest eigenvalue of the correlation matrix of the variables.

In areas such as program evaluation, it is typically the *first principal component* that an applied economist utilizes, or PC1. Suppose a researcher has assembled a set of survey items that she believes collectively measure anxiety about food insecurity: worry about running out of food, sleep disruption, feelings of inadequacy as a provider, shame in social situations involving food. If the first principal component gives high positive weights to most of these items, this is evidence that they share a common underlying signal. The component score can then serve as a summary measure of that latent construct. If PC1 accounts for, say, 45 percent of total variance across ten anxiety items, the researcher has reasonable grounds for concluding that there exists a coherent underlying dimension of food-related psychological distress that her questions are jointly capturing.

The technical foundations of PCA rest on the eigenvalue decomposition of the covariance matrix of the data. The eigenvector indicates the direction of the variance in K-dimensional space and the eigenvalue the strength of this internal correlation of the index items, indicating the share of total variance that component explains. Here PCA finds the best low-dimensional representation of high-dimensional data, preserving as much of the original variation as possible. PCA is most appropriate when the component variables are each understood as noisy, imperfect measures of a single underlying latent construct and where simple dimensionality reduction for index components is a key research goal.

In this setting, the variance-maximizing weights of the first principal component do genuine conceptual work that equal weighting cannot. The limitation of PCA is the mirror image of its strength: the weights it assigns are determined entirely by the variance structure of the data, not informed by theory.

3.5 Factor Analysis

Factor Analysis (FA) is the closest relative of PCA in the index construction toolkit, and the two are frequently conflated in applied work. This conflation is understandable as both psychometric techniques reduce the dimensionality of a set of correlated variables, and both produce weighted combinations that can serve as summary indices. The distinction, however, is conceptually fundamental and practically consequential.

On one hand, PCA is agnostic about the existence of an underlying latent variable driving a set of observable outcomes. At its core, PCA is a mathematical dimensionality reduction operation. Factor analysis, in contrast, posits that observed variables are generated by a smaller number of latent factors (plus item-specific error) and then estimates those factors from the data. Where PCA asks "What combination of these variables explains the most variance?", FA asks "What latent constructs would have produced these observed correlations?" While both tend to create *reflective* indices, factor analysis in the truest sense creates a latent construct index.

Formally, the factor analytic model expresses each observed variable x_{ik} as x_{ik} =

$\sum_{m=1}^{M} \lambda_{km} F_{im} + \epsilon_{ik}$, where $F_1, \ldots, F_M$ are the latent factors, λ_{km} are factor loadings indicating how strongly variable k is associated with factor m, and ϵ_{ik} is the unique variance specific to variable k for observation i that is not shared with any other variable. This last term reflects a crucial difference with PCA: factor analysis explicitly partitions the variance of each variable into *common variance* shared with the latent factors and *unique variance* that is variable-specific. PCA, in contrast, uses the total variance and makes no such partition, and thus incorporates noise unrelated to the latent factor into the index. In an index construction context this matters because unique variance is noise relative to the latent construct of interest; factor analysis discards it while PCA absorbs it.

FA then uses the loadings to construct weights assigned to each variable, most commonly through the *regression method*. These are essentially factor loadings adjusted by the correlation matrix $\mathbf{R}$ for correlations across variables. The weights for the index created by the first factor are thus $\mathbf{w} = \mathbf{R}^{-1}\boldsymbol{\lambda}_1 = [\text{Var}(\mathbf{z})]^{-1}\,\text{Cov}(\mathbf{z}, f_1)$, which yields the factor analytic index:

$$FA_i = \sum_{k=1}^{K} w_k^F z_{ik}$$

Factor analysis has two main variants: *Exploratory factor analysis* (EFA) makes no prior assumptions about which variables load onto which factors, allowing the loading structure to emerge from the data. It is most appropriate early in instrument development, when the researcher is uncertain how many latent dimensions her items reflect. *Confirmatory factor analysis* (CFA) instead specifies in advance which variables are hypothesized to load onto which factors and tests whether the data are consistent with that structure. CFA is the more rigorous tool for validating an established instrument in a new population, a common and important task in behavioral development economics applications, where psychological scales developed in high-income settings require validation before use.

Factor analysis is a psychometric tool, and psychometric instruments should meet both *reliability* and *validity* criteria, the former being a necessary but not sufficient condition for the latter. Reliability encompasses 1) retest reliability; 2) parallel test reliability; 3) reliability across examiners; and 4) internal consistency across index variables (Nunnally, 1978). Reliability is traditionally tested via Cronbach's alpha where $\alpha = \left(\frac{K\bar{c}}{\bar{v}+(K-1)\bar{c}}\right)$ and K is the number of items in the index, $\bar{c}$ is the average covariance between items, and $\bar{v}$ is the average variance of the items. Cronbach's alpha $\gtrsim$ 0.70 indicates that variables in an index are sufficiently correlated to measure the same underlying construct.

For an index to be *valid*, it must be *reliable*, but also satisfy other validity conditions: *face validity*, accurately capturing the construct it intends to measure; *content validity*, including all important facets of the construct; *criterion validity*, correctly predicting behavior; and *construct validity*, where the index truly measures the theoretical construct it is intended to measure, correlating strongly with other measures of the same construct, and not correlating strongly with other constructs (Nunnally, 1978; Furr, 2021). In cross-cultural contexts, reflective indices also require careful attention to local validity; instruments developed in western industrialized countries may not translate well to other global settings (van de Vijver and Leung, 1997).

This brings us to a common question in development research: what are the respective contexts for appropriately using PCA and CFA?

Psychologists generally prefer factor analysis for identifying latent constructs because (a) it is much more amenable to theory-building; (b) it screens out the influence of variation not associated with the underlying construct; and (c) it strips away the error unique to underlying individual variables, thus showing how much variation is actually driven by the latent factor. In general, factor analysis is preferred over principal components when it is important to identify a specific latent construct. And for applied researchers working with established psychological constructs, or seeking to validate an imported instrument into a new cultural context, factor analysis offers a depth of inferential leverage that PCA cannot match.

This being said, principal components tends to be more widely used in applied economics research across development, labor, and often health economics. Why? First, in many applications, indices are often constructed from variables that capture distinct but related dimensions of a broader concept, such as financial literacy, business practices, or household assets. In such cases, the assumption of a single underlying latent factor is often inappropriate or irrelevant. Using PC1 is often better suited to these hybrid constructs (Filmer and Pritchett (2001b); Vyas and Kumaranayake (2006)). This distinction reflects the broader difference between latent-variable modeling and data reduction emphasized in the measurement literature (Fabrigar et al., 1999).

Factor analysis also has quite a few moving parts, decision-making over the threshold for meaningful factor loadings, whether to use an orthogonal or oblique rotation and even sub-types of rotation within each of these broad categories. Although factor rotation is irrelevant when only a single factor is extracted, PCA does, nevertheless, leave somewhat less wiggle room for researchers, is more transparent, more easily replicable, and more easily interpretable.

A third reason is robustness. CFA, which tries to recover a latent covariance structure, performs well in large samples, but less dependably in small ones. The traditional sentiment in psychometric heuristics was that factor analysis required large samples, on the order of $n > 300$ (Comrey and Lee, 1992). Subsequent simulation studies indicated these CFA rules to be overly blunt (Marsh et al. (1998); MacCallum et al. (1999); Wolf et al. (2013)) because the stability of factor analysis depends on more than sample size. In cases where factor loadings and the number of indicators are relatively low, small sample sizes can create instability (Fabrigar et al. (1999); Wolf et al. (2013)), but where there are multiple indicators with high factor loadings, simulations have shown factor analysis to be stable for n as small as 100 (MacCallum et al., 1999). Thus in practice, there is no universal sample size cutoff, but where psychological and survey-based measures may be noisy, such as in some LMIC contexts, samples below several hundred observations may yield unstable factor solutions. And in these contexts, using PC1 is likely a more stable alternative. A good rule-of-thumb is to use PC1 in very small samples and to use PC1 as a robustness check on CFA for $n \lesssim 1000$.

All this said, when using a well-designed questionnaire whose items exhibit strong internal consistency (Cronbach's $\alpha > 0.70$ (Danner, 2016)) and a reasonably large sample, empirical differences between indices based on PC1 and those based on the first factor are often modest, particularly when a clear dominant factor underlies the items (Bikos, 2024; Mueller, 2025). PC1 will tend to assign heavy weights to the same index variables that the first CFA factor does. For this reason, principal components can serve as a solid alternative or robustness check even for psychological and social index construction in applied economic research, especially with modest or small sample sizes.

3.6 Constant Elasticity of Substitution (CES) Index

Sometimes we are unwilling to assume perfect substitutability between the components of a development index. Yet this assumption is implicitly imposed by any linear aggregation method such as straight scale, Kling, Anderson, principal component, and factor indices. An index based on a constant elasticity of substitution allows for that flexibility.

CES development indices allow researchers to aggregate multiple welfare dimensions while controlling the degree of substitutability among them. CES indices allow improvements in one dimension to compensate only partially for deficits in others, thereby embedding normative assumptions about the nature of trade-offs between different components of welfare. The general form is

$$\text{CES Index}_i = \left(\sum_{k=1}^{K} w_k x_{ik}^{\rho}\right)^{1/\rho}$$

where x_{ik} represents the level of development dimension k for individual or household i, w_k are weights satisfying $\sum_{k=1}^{K} w_k = 1$, and ρ determines the elasticity of substitution across dimensions.

Assuming individual index components are standardized, $\rho = 1$ replicates a Kling index where there is perfect substitutability between the components. For $\rho \in (0, 1)$ there is partial substitutability, with $\rho \to 0$ being the special Cobb-Douglas case.

3.6.1 Special CES Case: The U.N. Human Development Index

The United Nations Development Programme Human Development Index (HDI) has evolved into a special case of a CES index. Before 2010 the HDI used a linear average of metrics for health (H), education (E), and per capita income (I), or $\frac{1}{3}(H + E + I)$, which assumed perfect substitutability between the three components. In 2010 the HDI began to use the geometric mean $(H \cdot E \cdot I)^{\frac{1}{3}}$ in order to penalize uneven country development across the index components (United Nations Development Programme, 2010). The new(er) HDI thus assumes a special case of the Cobb-Douglas form with each weight equal to $\frac{1}{3}$, thereby recognizing the imperfect substitutability of health, education, and income in human development.

3.6.2 Special CES Case: The Rawlsian Index

A Rawlsian Index (varyingly referred to also as a Bottleneck, Leontief Aggregator, or Minimum Outcomes index) represents another case of a CES-utility-based index in which substitutability is zero between index components ($\rho = -\infty$). A Rawlsian Index defines welfare by its weakest component and can be re-written as

$$\text{Rawlsian Index}_i = \min(y_{i1}, y_{i2}, \ldots, y_{iK})$$

where $y_{i1}, \ldots, y_{iK}$ are (potentially standardized) scores across K index components.

The philosophical lineage runs through John Rawls: just as his difference principle evaluates social arrangements by the welfare of most deficient human being, a minimum outcomes index evaluates individual welfare by its most deficient human dimension. The index is most appropriate when welfare is genuinely non-compensatory across dimensional domains–when no amount of achievement in one dimension can offset severe deprivation in another. Such an index best represents settings characterized by a set or series of necessary but not sufficient welfare conditions.

The application to areas such as health is self-evident: People are generally only as healthy as their least healthy organ. A person who is strong, mobile, and cardiovascularly healthy but suffers from a malignant brain tumor is not, in any meaningful sense, healthy. Similarly, a road that is well-graded with perfectly painted traffic lines, yet has a single collapsed bridge is impassable. The Rawlsian Index may apply well to some water, sanitation, and hygiene (WASH) interventions. A WASH intervention resulting in clean water from a borehole, clean jerry cans, and clean drinking glasses, but that does not address bacteria accumulation in household water storage should yield a grade in a WASH index no higher than that of the most contaminated storage container. In each of these cases, the strong performance across the functional WASH components is irrelevant. Any index based on linear combinations of composite variables is inapplicable.

The principal limitation to a Rawlsian index is its sensitivity to measurement error in any single component, a vulnerability that makes careful attention to each constituent measure especially important. As a Rawlsian index represents the extreme form of complementarity, it is unable to capture any improvement in welfare stemming from an improvement in any but the binding composite variable.

3.7 Alkire–Foster Indices

Alkire and Foster (2011a) develop a general framework for multidimensional poverty measurement based on a dual-cutoff counting approach. The method separates the problem into an *identification* step (who is poor) and an *aggregation* step (how poor they are), and satisfies a set of desirable axioms for multidimensional poverty indices.

Consider a population of n individuals and K dimensions. Let x_{ik} denote individual i's achievement in dimension k, and ϕ_k the deprivation cutoff in that dimension. The deprivation matrix g_{ik} is defined by

$$g_{ik} = \begin{cases} 1 & \text{if } x_{ik} < \phi_k, \\ 0 & \text{otherwise.} \end{cases}$$

Each dimension has an associated weight $w_k > 0$ with $\sum_{k=1}^{K} w_k = 1$. The weighted deprivation score for individual i is

$$c_i = \sum_{k=1}^{K} w_k g_{ik}.$$

A second cutoff $\psi \in (0, 1]$ identifies the multidimensionally poor: individual i is poor if $c_i \geq \psi$. This dual-cutoff rule generalizes simple union/intersection rules and is central to the AF method (Alkire and Foster, 2011b).

Let q be the number of poor. The *headcount ratio* (incidence) is

$$H = \frac{q}{n},$$

and the *intensity* of poverty is the average deprivation score among the poor,

$$A = \frac{1}{q} \sum_{i:c_i \geq \psi} c_i.$$

The core AF measure, the *adjusted headcount ratio*, is

$$M_0 = H \times A = \frac{1}{n} \sum_{i:c_i \geq \psi} c_i,$$

which can be interpreted as the average weighted share of deprivations among the population (Alkire and Foster, 2011a).

Alkire and Foster show that M_0 is *decomposable* by subgroup and dimension, and that it satisfies key axioms such as dimensional monotonicity and poverty focus. They extend the framework to incorporate the depth and severity of deprivations via M_1 and M_2, which use censored poverty gaps and squared gaps, respectively, paralleling the Foster–Greer–Thorbecke family in the unidimensional case. The AF methodology underlies the global Multidimensional Poverty Index (MPI) used by UNDP and many national statistics offices United Nations Development Programme (2018).

It is correct to think of the AF methodology as a general framework for measuring multidimensional disadvantage, not just poverty headcounts. It has been applied to construct indices of multidimensional well-being and quality of life, child deprivation, women's empowerment and gender inequality, and multidimensional vulnerability or resilience (Alkire and Kovesdi, 2020; Tesfamariam et al., 2020; Alkire et al., 2013b; Jones and Tanner, 2013). In each case, researchers define dimensions, deprivation cutoffs, and weights; compute a weighted deprivation score; apply a dual cutoff (within-dimension and across-dimensions); and aggregate via an adjusted headcount-style measure. Thus, the AF index is best viewed as a flexible counting-based multidimensional index framework, with poverty measurement as its most visible application.

3.8 Which Index to Use?

The choice of index for different constructs and contexts is an important but understudied problem in applied research. Figure 2 presents a decision tree that seeks to facilitate this choice. The first consideration relates to the general type of index one seeks to create: *reflective*, *formative*, or a *hybrid* index that contains both reflective and formative components.

3.8.1 Reflective Indices of Latent Constructs

Branch A in the decision tree involves reflective indices, indices that are seeking to identify an underlying latent construct. This might be a well-defined mental health construct such as anxiety, resilience, or PTSD, or perhaps a somewhat less-well-defined construct such as aspirations, social capital, or empowerment. A first question rests over dimensionality. If an index is *a priori* multi-dimensional, then with a sufficiently high number of observations to ensure stability, multi-factor analysis is the appropriate option because of its ability to identify multiple latent factors within the data. With smaller sample sizes, multiple-component PCA offers greater stability and robustness, though when used in program evaluation, it is not strictly measuring effects on a latent factor.

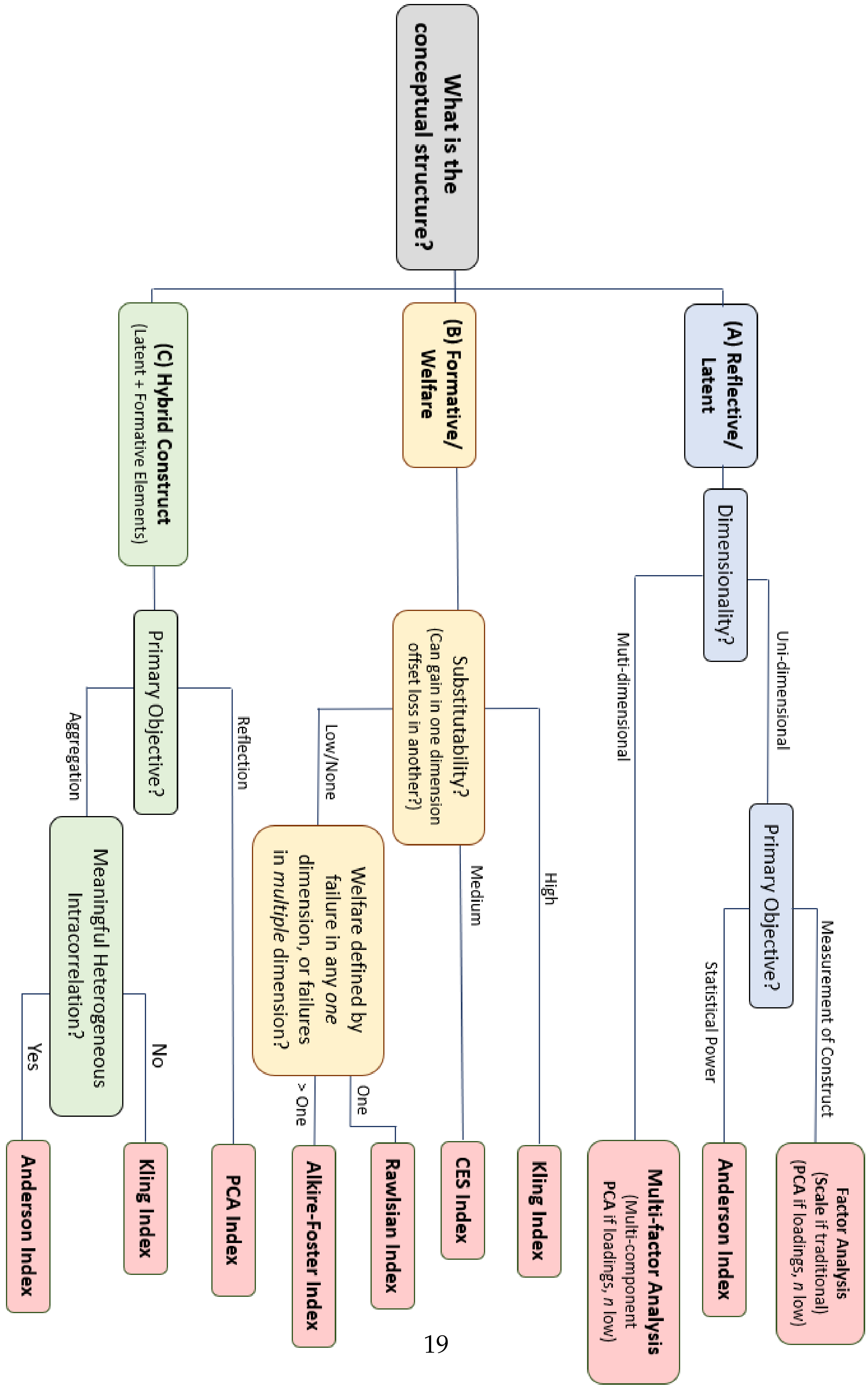
For uni-dimensional indices, the central distinction is whether the main objective is accurate *measurement* of a latent construct or *statistical power*.

The Anderson index is commonly used in development economics, especially in impact evaluation, when researchers often prioritize statistical power to identify treatment effects. As a GLS-based index, the Anderson index by definition maximizes statistical power. But when measuring outcomes with a theoretically strong latent construct, using the Anderson index solely to increase statistical power likely comes at the price of construct mis-measurement, and hence attenuation bias.

Factor analysis, unconcerned with minimizing variance in an index, places higher weights on variables that are most *strongly* correlated with the latent construct such that $\mathbf{w_F} \propto \mathbf{\Sigma}^{-1}\mathbf{\Lambda}$, where $\mathbf{w_F}$ is a $k \times 1$ vector of index weights, $\mathbf{\Sigma}^{-1}$ is the $k \times k$ inverse of the index variable covariance matrix, and Λ is a $k \times 1$ of factor loadings. This ensures that the resulting index is a scale-calibrated estimate of the underlying latent trait. In contrast, the GLS-based Anderson index, unconcerned with construct measurement, chooses weights to minimize index variance such that $\mathbf{w_A} \propto \mathbf{\Sigma}^{-1}\mathbf{1}$. By placing high weights on the variables in the index least correlated with other variables in the index, the Anderson index will often assign high weights to variables most *weakly* correlated with the underlying construct.

The differences between the two indices depend on heterogeneity in factor loadings on the construct. When CFA factor loadings across variables are close to identical, weights will converge between the two indices. However, when loadings are strongly heterogeneous, the indices are likely to produce quite different results in both point estimates and their standard errors.

Figure 2: Index Construction Decision Tree



As an example, suppose an index designed to measure child aspirations contains three variables, two strongly correlated with the underlying construct (e.g., aspired educational level and aspired profession) and one weakly correlated with it (say, perseverance or “grit" observed in an experimental task). CFA will assign larger index weights to the two strongly correlated variables and a smaller weight on the weakly correlated variable, while the Anderson index will place a relatively higher weight on the weakly correlated variable because it contains more independent information. Assume the aspired education and profession variables are more correlated with the underlying aspirations construct than the more weakly correlated grit observation. In this case Anderson index would yield an inferior aspirations construct measure, an empirical construct perhaps capturing more variability across ”grit" than aspirations.

By treating all variation as noise to be minimized, the Anderson index may inadvertently suppress the latent signal, resulting in attenuation bias and lower statistical power. CFA, by weighting according to the loading matrix $\mathbf{\Lambda}$, prioritizes signal recovery over variance minimization. Returning to our example, we should not be surprised when our intervention intended to elevate children’s aspirations yields a muted treatment effect when a very “gritty" Anderson index is regressed on the aspirations intervention.

Proposition 1 (see appendix for proof) shows that in general, use of a reflective index with a vector of variable weights incongruent to those comprising the latent factor will generally yield a biased treatment effect on the latent construct. This bias is equal to zero only in the serendipitous case that differences in the weighting vectors between the constructed index and the true latent factor index are orthogonal to the vector of estimated treatment effects on the underlying variables.

Because the Anderson index assigns weights $w \neq \theta$, even the power superiority of the Anderson index is not guaranteed. This is because the signal-to-noise contest for statistical power is a Godzilla-vs.-Kong-like battle between the optimized signal of psychometrics (factor analysis) and econometrically optimized noise (Anderson).

Which index will prevail in statistical power? This depends on whether the higher measurement precision of CFA outweighs the lower variance of the Anderson index. Suppose we wish to estimate the treatment effect of an intervention on a latent construct. Define $\beta = E[x_i|T = 1] - E[x_i|T = 0]$ as the $k \times 1$ vector of treatment effects on the components. We can use the non-centrality parameters to compare the "signal-to-noise" conditions that would favor a CFA index (F) over an Anderson index (A):

$$\frac{(w_F'\beta)^2}{w_F'\Sigma w_F} > \frac{(w_A'\beta)^2}{w_A'\Sigma w_A}$$

These occur when the "alignment" between index weights and treatment effects across the underlying index variables outweigh the loss in efficiency from moving away from a GLS-based minimum variance index, which will tend to occur at higher levels of loading heterogeneity. In extreme loading cases, the Anderson index places maximal weight on "quiet variables", variables which may be close to orthogonal to the underlying construct. Thus the Anderson index's pursuit of statistical efficiency can lead it into both conceptual and statistical power dilution.

Use of the Anderson index for latent factors is a possible path in Figure 2 but it is generally not the advisable path. Best practices favor the use of CFA over the Anderson index when measuring latent psychological or sociological phenomena with a clear construct except in clear cases in which statistical power is deemed a central priority and precise measurement of a construct is ancillary. In smaller sample sizes where CFA may be unstable, PC1 is likely to closely replicate outcomes of a first CFA factor and also forms a strong robustness check.

3.8.2 Formative Welfare Indices

As seen in Branch B in Figure 2, the primary question for choosing a formative welfare index is over the substitutability of index elements: Can a gain in one dimension of the index offset welfare from a loss in another dimension? If so, a Kling index is likely appropriate, such as might be the case in a wealth index, where a gain in one type of wealth may easily substitute for a loss in another area.

In practice, however, too many indices assume substitutability where it would not seem to exist. For example, many commonly used health-related indices, DALYs and QALYs, Burden-of-Disease Indices, and the SF-36, quizzically assume a very high level of substitutability between health outcomes when in practice substitutability across health outcomes is extremely low. Further increases in the health of a healthy organ does not usually substitute for the consequences of a failed one.

Another example of this is the World Health Organization Quality of Life indicator. In practice, WHOQOL domain scores are constructed as simple averages of item responses and are then linearly aggregated across domains. In the framework of a CES-based index $F = \left(\sum_{j=1}^{J} w_j x_j^{\rho}\right)^{1/\rho}$, it assumes $\rho = 1$ or perfect substitutability across domains such as physical health, social relationships, psychological health, and environmental quality, implying full compensability in the scoring rule. However, the conceptual framework underlying WHOQOL does not justify such substitutability, as domains such as physical health, social relationships, and environmental conditions are treated as distinct and essential aspects of well-being, such that substitutability is imperfect, i.e. $\rho < 1$.

In the extreme case when substitution is impossible across domains, either a Rawlsian or Alkire-Foster index may be appropriate. The contextual distinction here is whether poor welfare can be defined by failure in any *one* dimension (like a crucial human organ) or whether it is defined in terms of severity by failures across multiple dimensions within the index.

When aggregating across items in a formative index, the Kling index in some contexts presents an acceptable option despite its (often implausible) assumption of infinite substitutability across index components. This is particularly appropriate if the underlying outcomes themselves embody significant complementarities. For example, consider a dwelling quality index in which variables include the quality of walls, floor, and roof. Because rational homeowners would not infinitely improve wall quality at the expense of adding a leaky roof, a dwelling index need not duplicate these inherent complementarities. Because we would expect the quality of these complementary dwelling attributes to increase together, even indices that allow for high substitutability can acceptably measure differences in overall dwelling quality.

The formative-indicator literature emphasizes that when indicators are constitutive of a construct, they need not co-vary strongly and should not be evaluated using internal-consistency criteria such as Cronbach's α that are designed for reflective scales (Bollen and Lennox, 1991a; Jarvis et al., 2003a; Bollen and Diamantopoulos, 2017).

For formative indices, the Kling index offers an advantage in the context of missing data, following standard practice in applied development economics, where indices can be constructed creating standardized indices over all available non-missing components (Kling et al. (2007), Casey et al. (2012)), and that are consistent with best practices in the missing data literature (Allison, 2001). Treating missing observations is less straightforward in indices with unequal weights, where researchers are either tempted to replace missing data with variable means (which biases index variance downwards), replace missing data with more cumbersome imputation methods, or use listwise deletion (where one must then assume that data are missing at random).

3.8.3 Hybrid Indices

Many of the most widely used indices in development economics exhibit a hybrid structure as seen in Branch C of Figure 2, including indices for socio-economic status (SES), empowerment indices, dietary indices, food security, and multidimensional poverty indices. These either embody constructs that conceptually lie between statistical aggregations and a reflective latent factor, or they combine reflective measurement within individual domains with formative aggregation across distinct domains. A nested structure reflects the fact that while individual dimensions may be well-captured by latent constructs, the broader concept of welfare or development is inherently multidimensional perhaps with low substitutability between dimensions. Nested indices can be constructed using the decisions made for reflective indices in Branch A, and then aggregated using the decisions in Branch B to create the final index.

The key question for single-layer hybrid indices is whether in conception and context they seek to reflect a single latent factor or whether they are defined as aggregates of their underlying variables. This holds for both dependent and independent variable indices. If the primary objective is aggregation, then the next question is if there exists meaningful heterogeneous intracorrelation between index variables. A reasonable benchmark is whether any pairwise correlation coefficient between index variables exceeds 0.50 while at least one other lies below 0.20. If so, then an Anderson index is appropriate, and if not, then statistical power is unlikely to fall significantly using the more interpretable Kling index.

If a construct leans toward reflectivity but still possesses strong formative characteristics, such as SES or social inclusion, a PC1 index strikes a balance that assigns high weights to variables strongly correlated with the component (and hence an underlying construct) without having to assume that all of SES is caused by a single latent factor. Using PC1 is more appropriate for hybrid indices that are control variables, where measurement takes precedence over statistical power. Researchers may consider using an Anderson index for hybrid dependent variable indices to exploit their greater statistical power, but best practice would suggest at least a robustness check for measurement using PC1.

4 Appropriate Data Types for Index Construction

4.1 Data Categorization Bins

An index is only as sound as the alignment between its construction and the data foundation on which it is built. Even an elegant weighting scheme, when applied to inappropriate data, can produce an index that is precise in construction but mysterious in interpretation. What kinds of data are ideal—or even acceptable—for different types of indices? This question is central because the data appropriate for constructing formative welfare indices differ fundamentally from those required for reflective indices in psychology and behavioral economics.

Suppose we categorize index data into six broad bins: (1) *objective outcomes* (such as asset ownership, educational attainment, and physical health data such as height and weight); (2) *observed behaviors* (school attendance, occupational choice, sleeping under a bed net); (3) *reported behaviors* (study effort, business practices, preventative health practices); (4) *experimental behaviors* (decisions made by subjects within a controlled experiment); (5) *hypothetical behavioral responses* (how someone says they would behave in a given situation); and (6) *reported behavioral characteristics* (how a person self-describes to a researcher).

4.2 Data Pyramid for Formative Welfare Indices

We can create a pyramid based on the value of these types of data for creating formative welfare indices. Precise rankings are always arguable, but as seen in Figure 3, *objective outcomes* and *observed behaviors* are fundamentally important to the creation of welfare indices, while *hypothetical behavioral responses* and *reported behavioral characteristics* are valued little and rarely used. This preference is deeply rooted in the revealed preference tradition of economics (Samuelson, 1938), which holds that welfare is best inferred from actual choices made under real constraints rather than from stated intentions or self-descriptions. The concern is well-founded empirically: a large literature documents systematic divergence between stated and revealed preferences, particularly in low-income settings where social desirability bias and hypothetical bias are pronounced (Hausman, 2012; List and Gallet, 2001). Economists generally place little weight in subjective judgments and hypotheticals. Their traditional concern is economic welfare. The best indicators of this are variables that represent optimal behavior or outcomes realized in the context of economic constraints.

The theoretical foundations of formative index construction have been elaborated most clearly in the capabilities approach (Sen, 1985; Nussbaum, 2000; Robeyns, 2011; Alkire and Foster, 2011a). Sen (1985) argued that welfare indices should be anchored in achieved functionings or "achievements" — prioritizing what people are actually able to do and be — over preferences, hypotheticals, or subjective reports. This is reflected in the construction of the most influential formative welfare indices in development economics, including the asset indices pioneered by Filmer and Pritchett (2001b), which rely exclusively on objectively observed household characteristics, and the multidimensional poverty indices of Alkire and Foster (2011b), whose indicators are drawn from directly observable or verifiable outcomes.

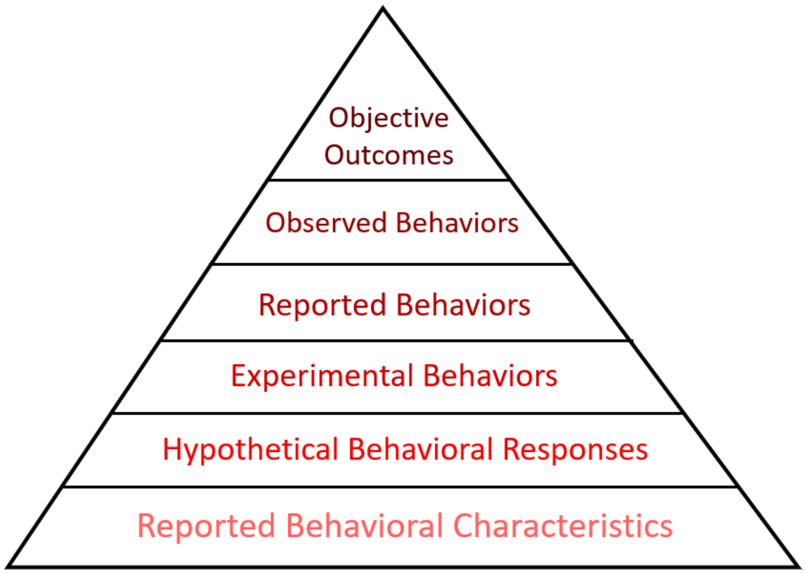


Figure 3: Hierarchy of Formative Index Data

The logic here is compelling: a formative index is one in which the components *create* the welfare construct, grounded in observable, measurable conditions rather than in self-report. In applied development work, this logic not only appears in multidimensional poverty measurement, but also in asset-based wealth indices, where researchers aggregate diverse material conditions into a welfare construct rather than treating diverse conditions as reflective symptoms of an underlying latent trait (Alkire and Foster, 2011a; Filmer and Pritchett, 2001a).

4.3 Data Pyramid for Reflective Construct Indices

Interestingly, the data pyramid for reflective indices is essentially turns the formative pyramid upside down. Figure 4 reflects the primary importance of reported behavioral characteristics and hypothetical behavioral responses. Why is this true?

First, this is because the logic of reflective measurement is fundamentally different than the logic behind economic welfare indices. When the object of interest is a latent psychological characteristic such as anxiety, self-efficacy, aspirations, or perceived agency, the indicators are modeled as manifestations of that latent trait rather than as the ingredients that compose it (Bollen and Lennox, 1991a; Jarvis et al., 2003a). This is why reported personal characteristics and hypothetical behavioral responses are often especially valuable. Self-report instruments are often better suited than behavioral or administrative measures for capturing internal states that are not directly observable, and item response theory formalizes this relationship by modeling responses to questionnaire items as functions of an underlying latent trait (Yang and Kao, 2014; Duckworth and Yeager, 2015). Hypothetical and vignette-based items can also be useful because they standardize context and hold many real-world constraints fixed, allowing the researcher to recover the construct of interest more directly (Aguinis and Bradley, 2014).

Critical to this understanding is that causality runs in the opposite direction between reflective and formative indices. In the formative index, it runs from observation to construct: statistically, food insecurity and low education *cause* one's categorization as "poor." In the reflective index it runs from the latent factor to subject articulation: anxiety *causes* a subject to state that a given event could cause them to cry.

A second reason is that while formative indices are created from *constrained* behavior, it is usually desirable to construct reflective indices from *unconstrained* behavior. Reported characteristics and hypothetical responses free participants from the behavioral constraints that are essential to recover welfare in formative indices. While constraints in traditional development economics inform development outcomes, in psychological measurement they often appear as statistical noise obscuring measurement of the underlying construct.

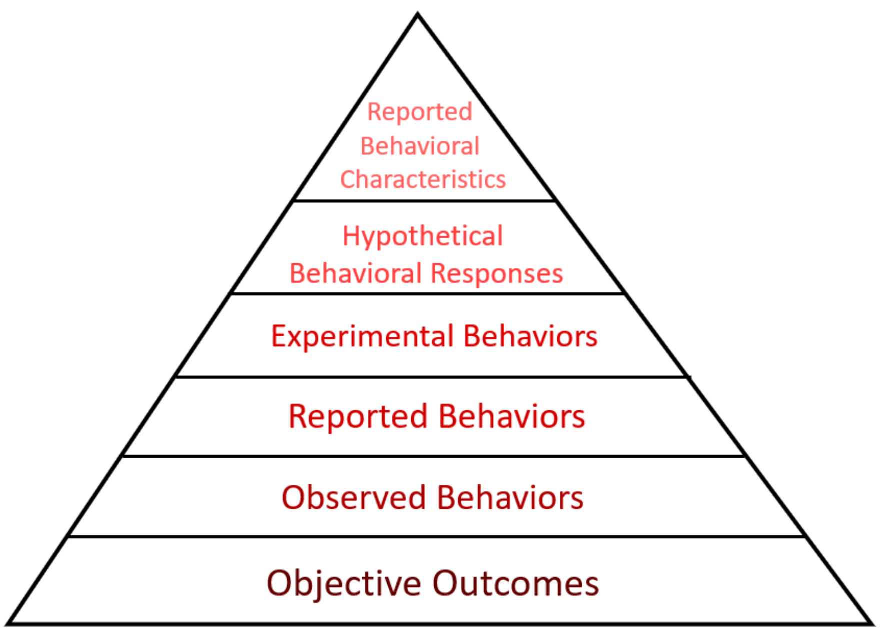


Figure 4: Hierarchy of Reflective Index Data

As a result, psychologists, behavioral economists, and others creating reflective indices value questions that recover latent constructs, even if hypothetical or self-described. This tradition has a long history in psychometrics, where validated instruments for anxiety (Spielberger, 1983), depression (Beck et al., 1961), and self-efficacy (Bandura, 1977) are constructed almost entirely from self-reported items precisely because the latent construct is understood to generate responses rather than to be constituted by them. This tradition has increasingly penetrated behavioral, health, and development economics as researchers have sought to measure psychological dimensions of poverty and well-being (Ridley et al., 2020b; Haushofer and Fehr, 2014).

This difference between formative and reflective index construction is not universally accepted, and some researchers have questioned whether it maps cleanly onto empirical practice. Bollen and Davis (2009) argue that the choice between formative and reflective measurement is frequently made on theoretical rather than empirical grounds, and that researchers often conflate the two in ways that generate invalid indices. A particularly common mistake in empirical economics is the construction of indices from self-reported items–Likert-scale measures of empowerment, agency, or mental health indicators using aggregation methods such as the Kling index, when the underlying measurement model is implicitly reflective and would generally call for principal components or factor analysis instead (Diamantopoulos et al., 2012). The reverse error is also documented: psychologists occasionally use factor-analytic methods on variables that would appear to create an index rather than related to outcomes that are created *by* the latent construct. This produces factors whose structure is an artifact of the method rather than a reflection of reality (Jarvis et al., 2003b).

In the middle of each pyramid sit types of data that have intermediate levels of usefulness in creating both reflective and formative indices: *reported behaviors* and *experimental behaviors*. Psychologists and behavioral economists often use these as imperfect but informative indicators of latent constructs — for example, self-reported frequencies of social interaction as proxies for sociability (Diener et al., 1985), or incentivized risk-taking tasks as indicators of underlying risk preferences (Holt and Laury, 2002). Early work by Fehr and Schmidt (1999) illustrates the aggregation of experimental behaviors into structured constructs where they estimate parameters of inequality aversion that summarize individual preferences across contexts. While not framed as a formative index, this approach combines multiple experimentally observed behaviors to construct a unified preference representation.

Research in behavioral economics has increasingly combined experimentally elicited measures of risk, time, and social preferences into joint constructs that characterize individual behavior. While again residing in a kind of middle ground between reflective and formative indices, this research often aggregates multiple experimentally derived behaviors into multidimensional preference profiles that function in similar ways to formative constructs: Gao et al. (2020) use experimental data to estimate individual-level preference parameters by combining outcome information across tasks, Harrison et al. (2022) elicit and analyze multiple preference parameters and beliefs as part of an integrated behavioral profile, and Falk et al. (2023) develop a validated instrument that jointly measures risk, time, and social preferences. These approaches illustrate how experimental behaviors can be aggregated into structured constructs, even if they remain confined to a single conceptual domain such as preferences rather than spanning multiple dimensions of welfare.

5 Variable Selection Criteria for Index Creation

5.1 Variable Selection Tradeoffs

How do we select the final set of variables for an index? The answer depends critically on whether the index is reflective or formative, as the underlying structure of these indices implies different selection criteria (Bollen and Lennox, 1991b).

Suppose we have a set $\mathcal{K} = \{x_1, x_2, \ldots, x_K\}$ of K candidate variables x_k, that each capture some aspect of the broader construct of interest. Why don't we simply throw the whole $\mathcal{K}$ in our index? There are several reasons to strive for parsimonious indices. The first is cost. Collecting data on borderline-useless index information carries an opportunity cost of collecting useful information. Moreover, every added question we include on a survey is a tax on participant patience and focus. (In an experiment to identify the costs of survey fatigue on data quality Jeong et al. (2023) find that an extra hour of surveying increased the probability of a respondent skipping a question by 10%–64% and reduced reported consumption by 25%.) Moreover, adding marginal "straggler" variables with a low signal-to-noise ratio reduces reliability within reflective indices. Parsimonious indices balance the need for information with these costs. Thus our aim should be to select a parsimonious subset of size $\tilde{K} \leq K$ for final index inclusion.

Variable selection for *formative* indices involves a tradeoff of coverage vs. redundancy. The tradeoff for *reflective* indices is one of measurement vs. noise. Because of their fundamental differences, variable selection for these different types of indices involves conceptually distinct tasks, which I present within a general framework tailored to empirical applied economics.

5.2 Variable Selection for Formative Indices

For formative indices, each variable added to an index potentially represents a distinct and irreplaceable dimension of the construct itself. Much must be left to the judgment of the index creator in this regard. Omitting a variable risks leaving part of the construct unaccounted for. Retaining variables capturing phenomena outside of the construct compromises the definition of the construct, and retaining redundant variables inflates the weight of the dimensions they share.[2] Selecting variables for formative indices involves a tradeoff between conceptual coverage and redundancy.

Canonical examples of formative indices include the UNDP's Human Development Index, with three domains: (1) *Health* (life expectancy); (2) *Education* (schooling completion); and (3) *Income* (per capita GDP) (United Nations Development Programme, 2010). More inclusive to other important human outcomes, the Global Human Flourishing index (VanderWeele, 2017), includes the following domains: (1) *Happiness and Life Satisfaction* (overall life satisfaction and positive affect), (2) *Mental and Physical Health* (self-rated physical health and mental well-being, including anxiety and depression), (3) *Meaning and Purpose* (a sense that life is worthwhile, with direction or calling), (4) *Character and Virtue* (moral traits such as honesty, integrity, generosity, and self-control), (5) *Close Social Relationships* (quality of relationships with family and friends, social support, and belonging), (6) *Financial and Material Stability* (ability to meet basic needs and maintain financial security), and (7) *Spiritual or Religious Well-Being* (religious participation, spiritual peace, or sense of connection beyond oneself, sometimes integrated into *Meaning and Purpose*).

Other work on human flourishing, such as Aten and Topping (2020) and Wydick et al. (2026) considers five human flourishing domains (*Physical Health, Economic/Education, Social Connectedness and Inclusion, Psychological Wellbeing, and Spiritual Wellbeing*) based on the compiled mission statements of faith-based non-government organizations (NGOs). However, formative indices are used ubiquitously in the development literature at more intermediate levels than overall human flourishing, to capture general outcomes in health (Kremer and Miguel, 2004; Dupas, 2011), education (Duflo, 2001; Anderson, 2008; Angrist et al., 2023), and microenterprise development (Banerjee et al., 2015; Rojas Valdes et al., 2022).

A strong criterion balances the inclusion of conceptually distinct domains against redundancy across variables. The key idea is that a well-constructed formative index should

[2] Redundancy is not limited only to formative indices; reflective indices also may contain redundant variables. Hall (1999) proposes an early machine-learning method for variable selection that embodies the idea that strong indices consist of variables that are highly correlated with an underlying latent factor, yet uncorrelated with each other.

span the relevant dimensions of the construct while avoiding duplication of closely related components. This yields a simple and transparent rule that complements the reflective case and is particularly useful in multidimensional welfare analysis.

5.2.1 A Simple Selection Algorithm for Formative Indices

Regularization methods such as the LASSO (least absolute shrinkage and selection operator), elastic net, and ridge regression, have been proposed for economic index selection in work such as Thompson and Talafha (2017) and Ye et al. (2022), but these have been mainly directed towards refining composite indices of economic activity and macroeconomic forecasting. In constructing formative indices in applied microeconomics, it is essential for a selection method to combine researcher intuition over necessary coverage of domains with a selection method that weeds out variables representing redundant domains.

Let each variable (or subindex) x_k be associated with a conceptual domain $d(k)$ with $k \in \{1, \ldots, K\}$, where K is the number of candidate domains and $\mathcal{K}$ is the set of those domains. The following proposes a simple LASSO-based algorithm for variable selection in formative indices:

1. Step 1: Standardize all variables in $\mathcal{K}$ (if not yet standardized).

2. Step 2: Choose some number of a "anchor" variables, $\mathcal{A} \subset \mathcal{K}$, perhaps 3-5 variables identified as *essential domains* within the index. These may be chosen subjectively or, without sufficient priors, chosen as the 3-5 variables showing the highest factor loadings on PC1 over the set $\mathcal{K}$. Using PC1, form the variable y_a from $\mathcal{A}$.

3. Step 3: Use cross-validation LASSO to select among the remaining m variables in $\mathcal{K}$ to complete the index, where using PC1 from Step 2 as the dependent variable LASSO selects $l \leq m$ among the (non-anchor) m variables for the index

$$\min_{\beta} \sum_{i=1}^{n} (y_a - X'\beta)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$$

based on the penalty parameter λ that is optimized through p-fold cross-validation.[3]

The set $\mathcal{L}$ of the L additional variables selected by LASSO will share some important characteristics. They will be strongly correlated with the $\mathcal{A}$ variables in the anchor. This

[3]This can be performed easily in R using the **glmnet** package and setting $\alpha = 1$ for LASSO and in Stata using the **lasso** command.

is important because generally there will exist high levels of correlation between variables in a formative index. (For example, author calculations show the average correlation between the three variables used in the 2025 Human Development Index to be 0.773.) But simultaneously there will be little statistical redundancy within the set of selected variables. LASSO will assign a zero weight to heavily redundant variables, those very strongly correlated with the selected variables.

The final set of variables used in the formative index will thus be $\mathcal{A} \cup \mathcal{L} = \tilde{K}$. Especially for formative indices, LASSO selection should serve as a guide rather than a rigid rule, and careful consideration should be given to the selection of anchor variables.

5.3 Variable Selection for Reflective Indices

Research since the early foundational work of Spearman (1904) and Cronbach (1951) in psychology has considered the tradeoffs involved in item selection for reflective indices, particularly the tension between internal consistency, construct validity, and parsimony. Early contributions such as Clark and Watson (1995) and Hall (1999) emphasize principles of scale development, including the importance of item homogeneity, content coverage, and the avoidance of redundancy. A second strand of work, including Fabrigar et al. (1999) and Marsh et al. (2013), focuses on the use of exploratory and confirmatory factor analysis to guide item selection, highlighting both statistical criteria for dimensionality and the risks of over-reliance on mechanical factor retention rules. More structurally oriented approaches, such as Hayduk and Littvay (2012), critique conventional latent variable modeling practices and stress the importance of theoretical coherence in specifying measurement models.

More recent work has revisited these foundations with a critical lens. Flake and Fried (2020) and Fried and Flake (2021) argue for greater transparency and rigor in construct validation, emphasizing that many commonly used scales lack clear conceptual grounding. Goretzko et al. (2021) revisits *exploratory* factor analysis as a flexible but often misapplied tool, while Sijtsma and Pfadt (2021) questions the dominance of reliability measures such as Cronbach's α, advocating for a broader view of measurement quality.

One thing is certain with constructing indices, and reflective indices in particular, an old adage rings true: less, in many cases, may be more. Consider our case of a one-factor model: $x_k = \lambda_j \theta + \varepsilon_k$, where θ is the latent construct, λ_k is how strongly item k loads on it, and ε_k is item-specific noise. Suppose we form a simple average index from K standardized variables: $I_K = \frac{1}{K}\sum_{k=1}^{K} x_k$. Dividing this into "signal" and "noise", this becomes $I_K = \left(\frac{1}{K}\sum_{k=1}^{K} \lambda_k\right)\theta + \frac{1}{K}\sum_{k=1}^{K} \varepsilon$.

Proposition 2 (proof given in the appendix) shows that adding a marginal variable to an index improves reliability only if it improves the signal-to-noise ratio of the index. If item $K+1$ has a weak loading λ_{K+1}, or large idiosyncratic noise, then the new index $I_{K+1} = \frac{K}{K+1} I_K + \frac{1}{K+1} x_{K+1}$ may be less correlated with the true latent construct θ, even though it includes more information. We can write the expression for the fraction of "signal" within an index as equal to $\rho_K^2 = \frac{\left(\sum_{j=1}^{K} \lambda_j\right)^2 \sigma_\theta^2}{\left(\sum_{j=1}^{K} \lambda_j\right)^2 \sigma_\theta^2 + \sum_{j=1}^{K} \sigma_{\varepsilon j}^2}$, and thus we should add the marginal variable to the index only if $\rho_{k+1}^2 > \rho_K^2$. Proposition 2 shows this will only hold for some factor loading on the marginal variable $\lambda_{K+1} > \lambda_{min}$. Adding noisy variables tends to contaminate reflective indices.

The following approach for variable selection in reflective indices is grounded in the logic of latent variable models. It generalizes to indices ultimately created using factor analysis, principal components analysis, or Kling index construction, for which impact measurement takes priority over construct measurement. The key idea is that a well-constructed reflective index should explain a large share of the variation in its constituent variables. At the same time, the inclusion of additional variables should be penalized to avoid unnecessary complexity and noise. The goal here is index simplicity and transparency for the applied researcher seeking a practical rule that balances signal extraction against parsimony, and a selection process that uses a framework familiar to those with backgrounds in econometrics.

While the LASSO is ideal for variable selection in formative indices, its close machine-learning cousin the *elastic net* is better suited for variable selection in reflective indices. The reason is that the goal with reflective indices is not strictly coverage, as it is with formative indices, but reliability. In formative index creation we seek unique information; in reflective index creation we seek pure information and internal consistency. Indeed, some degree of redundancy is desirable in a reflective index in that it averages out measurement error of the underlying latent construct.

Like the LASSO, the elastic net shrinks regression parameters, but it represents a linear weighting between LASSO (which selects variables as well as shrinks parameters) and ridge regression (which shrinks parameters, but does not select variables). Thus the elastic net, a weighted average of the two based on its parameter $\alpha \in [0, 1]$, is well-suited for selecting variables as a block that are highly correlated with the anchor and one another, while removing variables that contain little additional information.

5.3.1 A Simple Selection Algorithm for Reflective Indices

For reflective indices, every candidate variable in the set $\mathcal{K}$ of index candidates is assumed to be an imperfect signal of the same underlying latent construct — but not all signals are created equal. Some of these variables will be strongly correlated with the underlying latent construct, while the signal-to-noise ratio in others may be low, yielding low correlation. Below are steps to variable selection for reflective indices:

1. Step 1: Standardize all variables in $\mathcal{K}$ (if not yet standardized).

2. Step 2: Form the anchor y_a from the PC1 of *all* variables $x_k \in \mathcal{K}$.

3. Step 3: Using the balanced penalty ($\alpha = 0.5$), run the elastic net with all of the candidate variables in $\mathcal{K}$ as independent variables:

$$\min_{\beta} \sum_{i=1}^{n} (y_a - X_i'\beta)^2 + \frac{\lambda}{N}[\alpha \sum_{j=1}^{K} |\beta_j| + (1-\alpha) \sum_{j=1}^{K} |\beta_j^2|]$$

4. Step 4: Run a "stress test" by creating and analyzing a coefficient value path plot for all $x_k \in \mathcal{K}$ that shows coefficients on all x_k across different levels of the penalty λ.[4]

5. Step 5: Extract the "straggler" variables that elastic net sets to a zero coefficient at low levels of λ. The recommendation is to use a one-standard deviation in the mean-square error (MSE) threshold, and select the $\tilde{K}$ variables to be included in the index based on those variables that the elastic net includes at a one-standard deviation distance from the value of λ (in the direction of parsimony) that minimizes the MSE. Additionally and alternatively, one can prune weak variables that by the plot function are only retained at very low levels of the penalty lambda.

In constructing reflective indices, note that PC1 is taken from *all* K candidate variables created for the anchor, while the formative-index approach uses only a subset of the K variables. This is because for pruning variables from reflective indices we want to use the entire item pool in constructing the latent anchor, minimizing the influence of item-specific measurement error. By leveraging the shared variance of all candidate indicators, we extract a high-fidelity common signal that serves as a more robust benchmark for elastic net purification.

[4]This can be performed easily in R using the **glmnet** package and using the **plot** function and in Stata using the **elasticnet** command and the **coefpath** function with the option **legend(on)**.

Although it is purifying the index, reducing the number of variables to $\tilde{K} < K$ in a reflective index may reduce Cronbach's α, decreasing nominal reliability. However, Cronbach's α can be initially bloated by a large number of moderately correlated variables in a reflective index. Using a smaller number of higher-quality items to create a persistent core increases index validity, especially if these variables have been shown to be strong indicators of the latent construct in previous research.

The resulting $\tilde{K}$ variables chosen for the reflective index will tend to balance parsimony with construct measurement, eliminating potential candidate variables that are less correlated with the anchor while retaining those most correlated with the anchor. When there is special need to reduce the number of survey questions due to research costs or participant fatigue, researchers may want to restrict variables at an even more conservative cutoff. As with formative indices, researchers should use machine-learning variable selection of index variables as a benchmark and guide, with flexibility for contextual researcher judgment.

6 Example: A Microenterprise RCT in Three Countries

6.1 Background of Microenterprise Development Study

I will apply this framework for index creation to the development of indices used in Agasha et al. (2026), which reports results from three pilot studies we ran in Uganda, India, and Peru on cross-cut microenterprise interventions, one seeking to develop "hard skills" in the form of financial literacy, the second seeking to build "soft skills" in entrepreneurial aspirations.

The first component of the intervention was an entrepreneurial aspirations (EA) intervention based on Snyder (1994), Lybbert and Wydick (2018b), and Rojas Valdes et al. (2022). The EA intervention involved the creation of 30-minute documentary films on successful entrepreneurs in each country. These documentary films were shown to entrepreneurs shortly after the baseline survey. The films were accompanied by an entrepreneurial aspirations curriculum that lasted five weeks on aspirations, agency, and ability to visualize pathways out of poverty (via enterprise expansion). After five weeks a follow-up survey was taken to measure impacts on each of these three main areas that are taken from the positive psychology framework of Snyder (1994)'s *Psychology of Hope*. Questions from these surveys are shown in the appendix. These sub-indices were then used to create a summary "Aspirational Hope" index.

The second intervention was a financial literacy (FL) intervention that used a farm-finance game built on a mobile phone app to teach basic financial skills. These skills involved calculating the profitability of investments, understanding interest rates, understanding financial risk, borrowing and microcredit, and liquidity issues. Financial literacy topics were taught in conjunction with plays of the farm-finance game.

A series of questions (available in the appendix) were administered to construct five sub-indices related to financial numeracy, financial knowledge, financial confidence, financial competence, and financial behavior. Like the entrepreneurial aspirations intervention, these five indices were then used to create a summary index we called "Financial Literacy".

6.2 Choice of Indices

6.2.1 Financial Literacy

We used two different types of sub-indices for our financial literacy variables. We chose a standardized straight-scale index for our financial numeracy sub-index. As seen in Table A1 the Financial Numeracy index was composed of a series of elementary finance-related quiz questions, whose answers were either right or wrong. The quiz was strongly influenced by and drew from previous financial literacy questionnaires, but was essentially original in composition. Were it a well-known index, a straight scale could have been possible, but being original to our research, it made more sense to standardize the index for better interpretation of results.

The remaining financial literacy sub-indices (Financial Knowledge, Financial Behavior, Financial Competency, and Financial Confidence), are arguably clear examples of hybrid constructs: Each perhaps leans toward having a latent factor, but they also embody strong formative elements. Moreover, they are not psychological or psycho-social constructs *per se*, but rather qualities that would arguably comprise facets of the index. Thus creating an index from the first principal component (PC1) is a logical approach to index creation for these variables, where we are choosing high variation over precise measurement of a latent factor.

The Financial Literacy uber-index, comprised of the five sub-indices, we identify as a formative construct. Here we are *defining* financial literacy as being composed of these sub-indices; causality runs from the sub-indices to the aggregated index. While we could possibly use an Anderson in such a formulation, here we consciously choose to give each of these sub-indices equal weight for clean interpretation of the Financial Literacy index, and so we use a Kling index to construct the Financial Literacy index from the sub-indices.

6.2.2 Entrepreneurial Aspirations

Unlike the financial literacy sub-indices, we consider the entrepreneurial aspirations outcomes to be produced by latent psychological factors related to aspirations, agency, and visualization of pathways. Hence we obtain each of these sub-indices using the first factor from confirmatory factor analysis. Given the fact that these are psychological indices, to create Anderson indices for use as an impact variables in the RCT would have been an attempt to increase statistical power at the expense of construct mis-measurement.

As with the Financial Literacy index, the uber-index of Entrepreneurial Aspirations was created using a Kling Index. Why not use a second layer of CFA? Our reasoning here was that we desired to follow Snyder (1994) who essentially *defines* his hope construct as consisting of 1) goals, 2) agency, and 3) pathways, where his own formation of the concept appears highly formative. Thus wanting to assign equal weight to these sub-constructs, we again used the Kling index.

6.3 Types of Data Used in Index Construction

What was the reasoning behind the variables we chose for the respective indices? Most of the questions as seen in Table A1 fall into the category of reported behavioral characteristics, hypothetical behavioral responses, and reported behaviors–those lying at the top of the reflective index pyramid (and at the bottom of the formative index pyramid.)

The Financial Knowledge index is comprised of three reported behavioral characteristics and one reported behavior. Financial Behavior and Financial Competency all are formed by reported behaviors, and Financial Confidence by two reported behaviors, two hypothetical responses, and four reported behavioral characteristics.

The exception is Financial Numeracy, which is a series of answers on a basic finance quiz that constitute a set of observable behaviors. For this reason the Financial Numeracy index was created using a (standardized) straight scale, simply reflecting the number of correct answers in the quiz, where the straight scale implicitly gives equal weight to each question.

We drew extensively from the psychology literature in forming our survey questions for entrepreneurial aspirations. Most of our questions originated from validated survey instruments in which we adapted questions originally used in the United States and Europe to local contexts in the three countries in which we carried out the study. All survey items for Aspirations, Agency, and Visualization of Pathways were reported behavioral characteristics, often in the form of stated personal values.

6.4 Variable Selection and Pruning of Indices

Although the indices we constructed for the study were relatively sparse, we used the elastic net procedure in Section 4 to screen variables for each of the sub-indices. For screening criteria, we used both the criteria of a variable being dropped at $\alpha = 0.50$ and the outlier criteria based on the plot coefficient path, where variables retained only at very low and outlying levels of the penalty parameter λ were culled from the index.

All variables were retained in our indices with the exception of one item associated with the Visualization of Pathways index, participant responses to the question "I become discouraged easily when I encounter obstacles in making money." This question exhibited a very small loading (0.035) on its PCA index of all items, where the next lowest item of the remaining four loaded at 0.477, and was only retained by elastic net regularization at very low levels of λ. The Visualization of Pathways index was thus restricted to the remaining four high-loading items.

6.5 Impact Estimates Generated by Different Indices

We used the following ANCOVA specification to estimate average treatment effects (ATEs) across our three countries in our cross-cut RCT:

$$Y_{i1} = \alpha + \tau_1 FL_i + \tau_2 EA_i + \tau_3 (FL_i \times EA_i) + X_i' \gamma + \theta Y_{i0} + \epsilon_i \tag{1}$$

where Y_{i1} and Y_{i0} are endline and baseline outcomes, respectively, FL_i designates assignment to the financial literacy treatment, EA_i is assignment to the entrepreneurial aspirations treatment, X_i is a vector of personal controls and ϵ_i is the error term.

We present summary statistics and regression results in the appendix, and for a complete review of these results see Agasha et al. (2026). For our present purposes we use these results to focus on differences in estimated treatment effects resulting from the use of alternative indices. Results on the PC1-constructed financial literacy indices are presented in Tables A3 and A4. We focus on the EA intervention and its outcomes for space considerations and because they involve impacts on more clearly latent constructs.

Figure 5: Estimates of the ATE and Confidence Intervals Across Index Types: Factor Analysis, Principal Components, and Anderson Indices

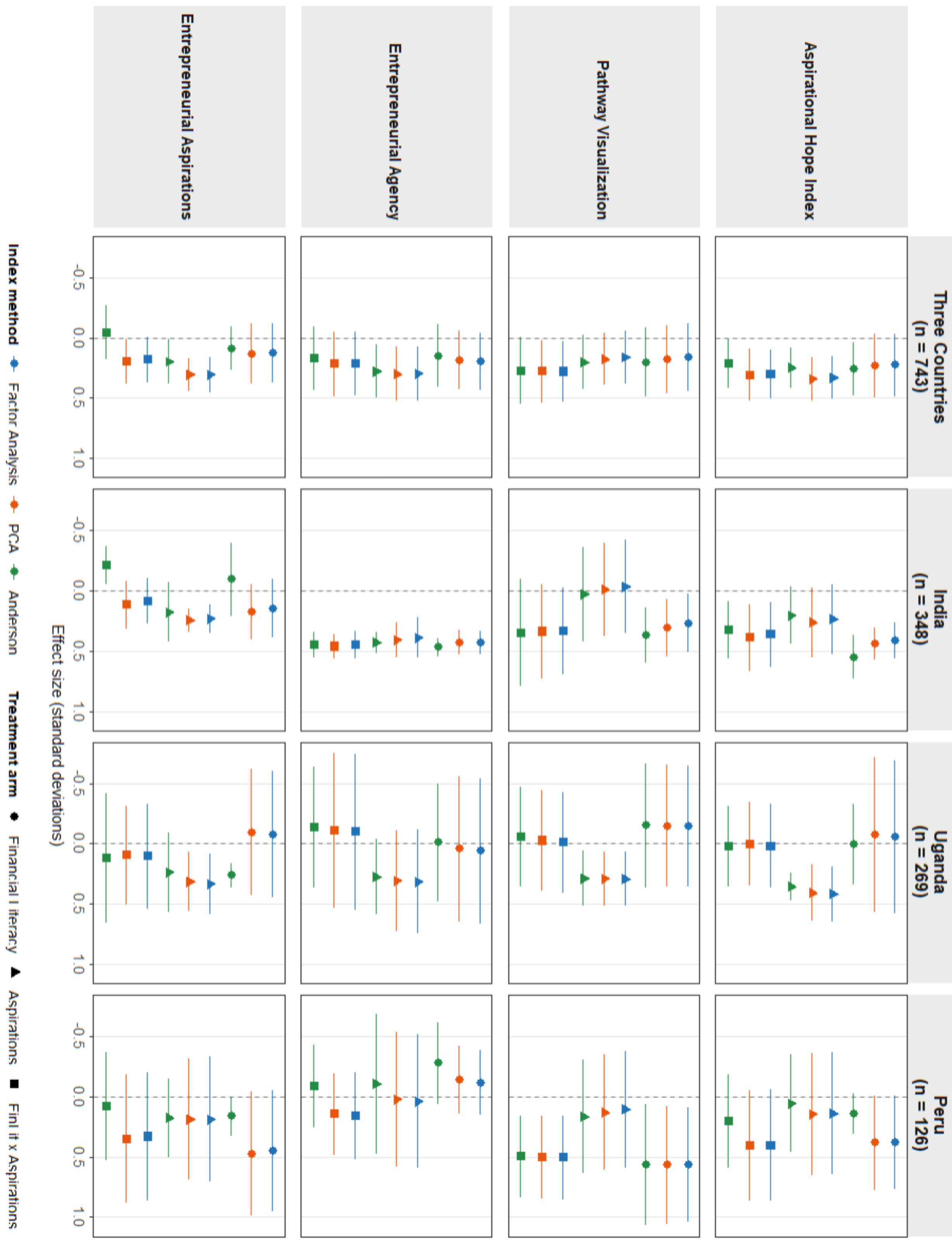


Figure 5 shows estimated ATEs on indices constructed using the first factor from CFA presented in Table A5, and the robustness checks using impacts on PC1-constructed indices and Anderson indices in Tables A6 and A7, respectively. Estimations are given for the three countries combined (n = 743), and on the smaller samples in each country, India (n = 348), Uganda (n = 269), and even smaller, Peru (n = 126). These single-country estimates are useful because they give us some indication on the relative performance of CFA, PCA, and Anderson indices across smaller sample sizes.

Figure 5 graphs the estimated ATEs across latent variable, across country, and across index construction type with impact data taken from Tables A5-A7. One can observe from the set of ATE point estimates and 95% confidence intervals in Figure 5 that while standard errors on point estimates appear to be generally lower using the Anderson indices, especially in the smaller country samples, point estimates tend to be lower as well. This is consistent with the positive and negatives of the Anderson index: greater estimation efficiency, but with downward attenuation bias in estimating impacts on latent constructs. But in the small sample in Peru, we can also see how CFA index standard errors can be quite large relative to Anderson, even in the aggregated Aspirational Hope index.

CFA is designed for measurement, not stastical power. Because both CFA (and its close relative, PCA) assign heavy weights to variables correlated with the underlying construct, we can view deviations from the point estimates of CFA in terms of bias.

Table 1: Tallies Across Index Methods: CFA, PCA, and Anderson

Tally Metric	**CFA**	**PCA**	**Anderson**
Lowest standard error	8	6	18
Highest standard error	14	7	11
Highest point estimate	13	12	7
Lowest point estimate	6	2	24
Highest $\|t\|$-statistic	9	12	11
Lowest $\|t\|$-statistic	10	4	18

Each cell reports the number of times (out of 32) that index method won the criterion, comparing the same outcome-arm-country cell across the Tables A5, A6, and A7.

Table 1 shows tallies across index methods between CFA, PCA, and Anderson indices across the 32 estimations in Tables A5-A7 in which the Entrepreneurial Aspirations treatment was assigned. The table tallies the number of times that each type of index displayed the lowest and highest standard errors across the three otherwise identical estimations, lowest and highest standard point estimates of the ATE, and lowest and highest t-statistics. Clear from these tallies is that while the Anderson index is far more likely to have the lowest standard error, it is also far more likely to have the lowest point estimate among the three (and the least likely to have the highest point estimate.) This is again consistent with Proposition 1 and the presence of attenuation bias in the Anderson index when it is used to represent latent constructs.

Although the Anderson index tallies the largest number of lowest-estimated standard errors, attenuation bias in its point estimates causes its t-statistics to suffer: it most often registers the lowest t-statistics among the three indices. While the Anderson index is able to reject null hypotheses of no impact for at least the 90% level from the EA and AE + FL interventions 12 times, both CFA and PCA are able to reject null hypotheses 14 times. As a result, use of the Anderson index to increase statistical power is at best questionable when estimating causal effects on latent constructs.

7 Conclusion

In recent years there has been a dramatic increase in the use of abstract constructs to measure human welfare. We typically measure these abstract constructs either through proxy variables or the creation of indices, constructs that are intended to capture summaries of welfare in a given domain or to embody an underlying latent phenomena or trait. This paper is intended to serve as a guide in defining the types of indices that can be used in applied research, the choice of which index to use in a given context, the choice of types of data to use in creating different types of indices, and the choice of what variables to include within an index. What follows are three summary points from this guide:

1.) **The choice between the commonly used economic indices (particularly Kling and Anderson indices) and the psychometric indices (particularly CFA and PCA) is often a battle between statistical power and construct measurement.** When construct measurement is primary, CFA should be given primary consideration. For hybrid indices, PCA is often a strong option that, while making no claim for precise measurement of a latent factor, generates sample-robust indices that like CFA, assigns heavy weights to strongly correlated variables. Anderson indices are built to increase statistical power, but may actually reduce it through attenuation bias.

2.) **While objective outcomes and observed behaviors are most appropriate for formative economic welfare indices, reported behavioral characteristics and hypothetical behavioral responses are appropriate in development and behavioral economics and an array of other applied fields.** While many empirical economists have shied away from these types of data, it is important to understand how they can recover latent phenomena better than the kind of observed behavior and outcome data typically used in economics. Reported behaviors and experimental behaviors often represent more versatile data, used across an array of different kinds of indices.

3.) **Indices can be pruned through well-understood machine-learning methods.** The LASSO is particularly suited to prune variables from formative welfare indices, the elastic net for pruning reflective indices. Moreover, these pruning tools need not be used in the context of hard rules, but as a flexible guide to help create parsimonious indices that reduce data collection costs, participant fatigue, and that are more easily interpretable as formed by a tractable number of variables.

An important message here is that economists should remain comfortable using the index tools common to economics (Kling, Anderson, CES, and Alkire-Foster indices) to create formative indices, but should also understand the measurement advantages of the psychometric tools for creating reflective indices (PCA and Factor Analysis). These tools were developed in psychology to address the challenges of measuring abstract psychological and psycho-social phenomena, and they do this better than the traditional economic indices.

When measuring impacts on latent constructs, best research practices will avoid substituting index tools designed for maximizing statistical power over those designed for optimal construct measurement. The indices more commonly employed in economics are often poor substitutes, and indices that poorly measure latent constructs may actually yield lower statistical power than those that measure latent constructs more precisely. Measure well, and the statistics often take care of themselves.

References

Agasha, E., A. Hobbs, A. Shaji, and B. Wydick (2026). Financial literacy and entrepreneurial aspirations: Results from pilot experiments in india, peru, and uganda. Working Paper #251, Center for Effective Global Action (CEGA), University of California at Berkeley. 34, 37

Aguinis, H. and K. J. Bradley (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods 17*(4), 351–371. 26

Alkire, S. and J. Foster (2011a). Counting and multidimensional poverty measurement. *Journal of Public Economics 95*(7–8), 476–487. 16, 17, 24, 25

Alkire, S. and J. Foster (2011b). Counting and multidimensional poverty measurement. *Journal of Public Economics 95*(7–8), 476–487. 16, 24

Alkire, S. and F. Kovesdi (2020). A bird's-eye view of well-being: Exploring a multidimensional measure for the united kingdom. *Human Development Report Office Research Paper*. Uses the Alkire–Foster method to construct a multidimensional well-being index. 17

Alkire, S., R. Meinzen-Dick, A. Peterman, A. Quisumbing, G. Seymour, and A. Vaz (2013a). The women's empowerment in agriculture index. *World Development 52*, 71–91. 6

Alkire, S., R. Meinzen-Dick, A. Peterman, A. Quisumbing, G. Seymour, and A. Vaz (2013b). The women's empowerment in agriculture index. *World Development 52*, 71–91. Constructs a women's empowerment index using the Alkire–Foster methodology. 17

Allison, P. D. (2001). *Missing Data*. Sage Publications. 22

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association 103*(484), 1481–1495. 2, 9, 29

Angrist, N., P. Bergman, C. Brewster, and M. Matsheng (2023). Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in botswana. *Nature Human Behaviour 7*, 1063–1072. 29

Aten, J. D. and S. Topping (2020). *Spiritual First Aid: Evidence-Informed Practical Guidance for Supporting People in Crisis*. Wheaton, IL: Humanitarian Disaster Institute. 29

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review 103*(6), 2121–2168. 1

Baird, S., R. Panlilio, J. Seager, S. Smith, and B. Wydick (2022). Identifying psychological trauma among syrian refugee children for early intervention: Analyzing digitized drawings using machine learning. *Journal of Development Economics 156*, 102822. 2

Bandiera, O., R. Burgess, N. Das, S. Gulesci, I. Rasul, and M. Sulaiman (2020). Women's empowerment in action: Evidence from a randomized control trial in africa. *American Economic Journal: Applied Economics*. 6

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review 84*(2), 191–215. 27

Banerjee, A. and E. Duflo (2013). Poor economics: A radical rethinking of the way to fight global poverty. *Society and Economy 35*(4), 573–587. 1

Banerjee, A. V., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science 348*(6236), 1260799. 29

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh (1961). An inventory for measuring depression. *Archives of General Psychiatry 4*(6), 561–571. 27

Bernard, T., S. Dercon, K. Orkin, G. Schinaia, and A. Seyoum Taffesse (2026). The future in mind: Aspirations and long-term outcomes in rural ethiopia. *Quarterly Journal of Economics 141*(1), —. 3

Bernard, T., S. Dercon, K. Orkin, and A. Seyoum Taffesse (2014). The future in mind: Aspirations and forward-looking behaviour in rural ethiopia. *Economic Journal*. 2

Bikos, L. H. (2024). Chapter 8: Principal components analysis. https://lhbikos.github.io/ReC_Psychometrics/PCA.html. Psychometrics course notes, accessed March 10, 2026. 14

Blattman, C. and J. Annan (2010). The consequences of child soldiering. *Review of Economics and Statistics 92*, 882–898. 2

Blattman, C., J. Jamison, and M. Sherman (2017). Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in liberia. *American Economic Review 107*(4), 1165–1206. 3

Bloom, N., R. Sadun, and J. Van Reenen (2020). Management as a technology? *Journal of Political Economy 128*(7), 2610–2668. 1

Bollen, K. A. and W. R. Davis (2009). Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling 16*(3), 498–522. 27

Bollen, K. A. and A. Diamantopoulos (2017). Notes on measurement theory for causal-formative indicators: A reply to hardin. *Psychological Methods 22*(3), 581–589. 22

Bollen, K. A. and R. Lennox (1991a). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin 110*(2), 305–314. 22, 25

Bollen, K. A. and R. Lennox (1991b). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin 110*(2), 305–314. 28

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press. 4, 6

Brazier, J. and D. Rowen (2021). The use of mapping methods to estimate health state utility values. *Pharmacoeconomics 39*(6), 615–631. 1

Casey, K., R. Glennerster, and E. Miguel (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics 127*(4), 1755–1812. 22

Chetty, R., N. Hendren, and L. F. Katz (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *Quarterly Journal of Economics 131*(2), 855–902. 2

Chinoy, S., N. Nunn, S. Sequeira, and S. Stantcheva (2024). Zero-sum thinking and the roots of u.s. political differences. Technical report. AEA Papers & Proceedings / Working Paper version widely circulated. 1

Clark, L. A. and D. Watson (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment 7*(3), 309–319. 31

Comrey, A. L. and H. B. Lee (1992). *A first course in factor analysis*. Psychology press. 14

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*(3), 297–334. 31

Cronbach, L. J. and P. E. Meehl (1955). Construct validity in psychological tests. *Psychological bulletin 52*(4), 281. 4

Dalton, P., S. Ghosal, and A. Mani (2016). Poverty and aspirations failure. *Economic Journal 126*, 165–188. 2

Danner, D. (2016). Reliability – the precision of a measurement. Gesis survey guidelines, GESIS – Leibniz Institute for the Social Sciences. Accessed March 10, 2026. 14

Deming, D. J. and L. B. Kahn (2021). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics 39*(S1), S337–S369. 1

Diamantopoulos, A., M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science 40*(3), 434–449. 27

Diener, E., R. A. Emmons, R. J. Larsen, and S. Griffin (1985). The satisfaction with life scale. *Journal of Personality Assessment 49*(1), 71–75. 27

Duckworth, A. L. and D. S. Yeager (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher 44*(4), 237–251. 26

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review 91*(4), 795–813. 29

Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics 3*, 425–449. 29

Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly 44*(2), 350–383. 1

Fabrigar, L. R., D. T. Wegener, R. C. MacCallum, and E. J. Strahan (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods 4*(3), 272–299. 13, 14, 31

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950. 28

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868. 27

Filmer, D. and L. Pritchett (2001a). Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of india. *Demography 38*(1), 115–132. 2, 25

Filmer, D. and L. H. Pritchett (2001b). Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. *Demography 38*(1), 115–132. 13, 24

Flake, J. K. and E. I. Fried (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science 3*(4), 456–465. 31

Fried, E. I. and J. K. Flake (2021). Measurement matters. *Association for Psychological Science Observer 34*(1), 28–29. 31

Furr, R. M. (2021). *Psychometrics: an introduction*. SAGE publications. 4, 13

Galor, O. and O. Ozak (2016). The agricultural origins of time preference. *American Economic Review 106*(10), 3064–3103. 1

Gao, X. S., G. W. Harrison, and R. Tchernis (2020). Estimating risk preferences for individuals: A bayesian approach. Technical report, Center for the Economic Analysis of Risk (CEAR), Georgia State University. 28

Genicot, G. and D. Ray (2017). Aspirations and inequality. *Econometrica 85*, 489–519. 2

Goretzko, D., T. T. H. Pham, and M. Bühner (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology 40*(7), 3510–3521. 31

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph. D. thesis, The University of Waikato. 29, 31

Harrison, G. W., M. I. Lau, and E. E. Rutström (2022). Subjective beliefs and economic preferences during the covid-19 pandemic. *Experimental Economics 25*(3), 795–823. 28

Haushofer, J. and E. Fehr (2014). On the psychology of poverty. *Science 344*(6186), 862–867. 27

Haushofer, J. and J. Shapiro (2016). The short-term impact of unconditional cash transfers to the poor: Experimental evidence from kenya. *Quarterly Journal of Economics 131*(4), 1973–2042. 1, 2

Hausman, J. (2012). Contingent valuation: From dubious to hopeless. *Journal of Economic Perspectives 26*(4), 43–56. 24

Hayduk, L. A. and L. Littvay (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology 12*(159), 1–17. 31

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655. 27

Jansen, A., R. Wang, P. Behrens, and R. Hoekstra (2024). Beyond gdp: a review and conceptual framework for measuring sustainable and inclusive wellbeing. *The Lancet Planetary Health 8*(9), e695–e705. 1

Jarvis, C. B., S. B. MacKenzie, and P. M. Podsakoff (2003a). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research 30*(2), 199–218. 22, 25

Jarvis, C. B., S. B. MacKenzie, and P. M. Podsakoff (2003b). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research 30*(2), 199–218. 27

Jeong, D., S. Aggarwal, J. Robinson, N. Kumar, A. Spearot, and D. S. Park (2023). Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics 161*, 102992. 28

Jones, L. and T. Tanner (2013). A multidimensional approach to measuring resilience. https://policy-practice.oxfam.org/resources/a-multidimensional-approach-to-measuring-resilience-302641/. Proposes a resilience index using the Alkire–Foster counting method. 17

Kabeer, N. (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and Change 30*, 435–464. 6

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119. 2, 8, 22

Kremer, M. and E. Miguel (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217. 29

List, J. A. and C. A. Gallet (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics 20*(3), 241–254. 24

Lybbert, T. J. and B. Wydick (2018a). Poverty, aspirations, and the economics of hope. *Economic Development and Cultural Change 66*(4), 709–753. 2, 3

Lybbert, T. J. and B. Wydick (2018b). Poverty, aspirations, and the economics of hope. *Economic Development and Cultural Change 66*(4), 709–753. 34

MacCallum, R. C., K. F. Widaman, S. Zhang, and S. Hong (1999). Sample size in factor analysis. *Psychological methods 4*(1), 84. 14

Marsh, H. W., K.-T. Hau, J. R. Balla, and D. Grayson (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate behavioral research 33*(2), 181–220. 14

Marsh, H. W., O. Lüdtke, B. Nagengast, A. J. S. Morin, and M. Von Davier (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right — camouflaging misspecification with item parcels in CFA models. *Psychological Methods 18*(3), 257–284. 31

Mueller, S. T. (2025). Exploratory factor analysis. https://pages.mtu.edu/~shanem/psy5220/daily/Day21/exploratoryfactoranalysis.html. Course materials, accessed March 10, 2026. 14

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill. 12, 13

Nussbaum, M. C. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press. 24

Olsson, O. and D. A. Hibbs (2005). Biogeography and long-run economic development. *European Economic Review 49*(4), 909–938. 1

Ridley, M., G. Rao, F. Schilbach, and V. Patel (2020a). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science 370*(6522). 2

Ridley, M., G. Rao, F. Schilbach, and V. Patel (2020b). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science 370*(6522), eaay0214. 27

Riley, E. (2024). Role models in movies: The impact of queen of katwe on students' educational attainment. *Review of Economics and Statistics 106*(2), 334–351. 3

Robeyns, I. (2011). The capability approach. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Winter 2020 Edition. 24

Rojas Valdes, R. I., B. Wydick, and T. J. Lybbert (2022). Can hope elevate microfinance? evidence from oaxaca, mexico. *Oxford Economic Papers 74*(1), 236–264. 3, 6, 29, 34

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica 5*(17), 61–71. 24

Sen, A. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland. 24

Sijtsma, K. and J. M. Pfadt (2021). Part II: On the use, the misuse, and the very limited usefulness of cronbach's alpha: Discussing issues about alpha. *Psychometrika 86*(4), 843–860. 31

Snyder, C. R. (1994). *The Psychology of Hope: You Can Get There from Here*. New York: Free Press. 34, 36

Spearman, C. (1904). "general intelligence," objectively determined and measured. *American Journal of Psychology 15*(2), 201–292. 31

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press. 27

Stiglitz, J. E. (2020). Measuring what matters. *Scientific American 323*(2), 24–31. 1

Tesfamariam, S. et al. (2020). Decomposing multidimensional child poverty and its drivers in the mouhoun region of burkina faso. *BMC Public Health 20*(135), 1–15. Applies the Alkire–Foster methodology to measure and decompose multidimensional child poverty. 17

Thompson, E. and A. M. Talafha (2017). Forecasting a composite indicator of economic activity in ghana: A comparison of data science methods. *Journal of Statistical and Econometric Methods 6*(4), 1–24. 30

United Nations Development Programme (2010). *Human Development Report 2010: The Real Wealth of Nations — Pathways to Human Development*. New York: Palgrave Macmillan. 15, 29

United Nations Development Programme (2018). Assessing multidimensional poverty – one index at the global level. https://hdr.undp.org/content/assessing-multidimensional-poverty-one-index-global-level. Human Development Report Office blog post: "Methodologically, the global MPI uses the Alkire–Foster method of multidimensional poverty measurement.". 17

van de Vijver, F. and K. Leung (1997). Methods and data analysis for cross-cultural research. *Cross-Cultural Research and Methodology Series 16*. 13

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proceedings of the National Academy of Sciences 114*(31), 8148–8156. 29

Vásquez-Lavín, F. et al. (2021). Estimating discount rates for environmental goods: The role of payment structures in stated preference data. *Journal of Environmental Economics and Management 107*, 102429. 1

Voors, M., E. Nillesen, P. Verwimp, E. Bulte, R. Lensink, and D. Van Soest (2012). Violent conflict and behavior: A field experiment in burundi. *American Economic Review 102*, 941–964. 2

Vyas, S. and L. Kumaranayake (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning 21*(6), 459–468. 6, 13

Wolf, E. J., K. M. Harrington, S. L. Clark, and M. W. Miller (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement 73*(6), 913–934. 14

Wydick, B., G. Camacho, and P. Piraino (2026). Hope walks: The impact of clubfoot treatment on human flourishing in ethiopia. *Health Economics 35*(1), 25–35. 29

Yang, F. M. and S. T. Kao (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry 26*(3), 171–177. 26

Ye, Y., R. Chi, Y.-H. Shao, C.-N. Li, and X. Hua (2022). Indicator selection of index construction by adaptive lasso with a generic ϵ-insensitive loss. *Computational Economics 60*, 971–990. 30

Ziegler, A. (2021). New ecological paradigm meets behavioral economics: On the relationship between environmental values and economic preferences. *Journal of Environmental Economics and Management 107*, 102516. 1

Zrubka, Z. et al. (2022). Predicting eq-5d-3l index scores from survey and clinical data: A comparison of machine learning and regression methods. *Journal of Health Economics 82*, 102579. 1

Appendix

Proofs of Propositions:

Proposition 1: Use of a reflective index with outcome variable weights incongruent to those comprising the latent factor will generally yield biased estimates of treatment effects on the latent factor.

Proof: Let the true latent construct be

$$y_i^* = \boldsymbol{\theta}'\mathbf{x}_i,$$

where $\boldsymbol{\theta}$ is a $K \times 1$ vector of factor scoring coefficients and $\mathbf{x}_i$ is a $K \times 1$ vector of outcome variables comprising the index. Suppose the researcher constructs the index

$$\tilde{y}_i = \mathbf{w}'\mathbf{x}_i,$$

where $\mathbf{w}$ is a $K \times 1$ vector of assigned weights, with $\mathbf{w} \neq \boldsymbol{\theta}$. Assume treatment is randomly assigned. Let

$$\boldsymbol{\beta} = E[\mathbf{x}_i \mid D_i = 1] - E[\mathbf{x}_i \mid D_i = 0],$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of treatment effects on the K index variables $x_1, x_2, \ldots, x_K$.
The average treatment effect on the true latent construct is

$$\tau^* = E[y_i^* \mid D_i = 1] - E[y_i^* \mid D_i = 0].$$

Substituting $y_i^* = \boldsymbol{\theta}'\mathbf{x}_i$ gives

$$\tau^* = \boldsymbol{\theta}' \left(E[\mathbf{x}_i \mid D_i = 1] - E[\mathbf{x}_i \mid D_i = 0]\right) = \boldsymbol{\theta}'\boldsymbol{\beta}.$$

By contrast, the treatment effect on the constructed index is

$$\tau_w = E[\tilde{y}_i \mid D_i = 1] - E[\tilde{y}_i \mid D_i = 0].$$

Substituting $\tilde{y}_i = \mathbf{w}'\mathbf{x}_i$ gives

$$\tau_w = \mathbf{w}' \left(E[\mathbf{x}_i \mid D_i = 1] - E[\mathbf{x}_i \mid D_i = 0]\right) = \mathbf{w}'\boldsymbol{\beta}.$$

Thus, the difference between the treatment effect on the constructed index and the treatment effect on the true latent construct is

$$\tau_w - \tau^* = \mathbf{w}'\boldsymbol{\beta} - \boldsymbol{\theta}'\boldsymbol{\beta} = (\mathbf{w} - \boldsymbol{\theta})'\boldsymbol{\beta}.$$

Therefore, the constructed index yields the same treatment effect as the true latent construct if and only if $(\mathbf{w} - \boldsymbol{\theta})'\boldsymbol{\beta} = 0$. Otherwise, $\tau_w \neq \tau^*$.

Proposition 2: *A marginal index variable $K+1$ reduces the reliability of a reflective index if its factor loading is some $\lambda_{K+1} < \lambda_{\min}$.*

Proof: The condition under which adding the marginal variable $K+1$ makes the index less reliable is $\rho^2_{K+1} < \rho^2_K$, where

$$\rho^2_K = \frac{\left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_\theta}{\left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_\theta + \sum_{j=1}^K \sigma^2_{\varepsilon j}}$$

and

$$\rho^2_{K+1} = \frac{\left(\sum_{j=1}^K \lambda_j + \lambda_{K+1}\right)^2 \sigma^2_\theta}{\left(\sum_{j=1}^K \lambda_j + \lambda_{K+1}\right)^2 \sigma^2_\theta + \sum_{j=1}^K \sigma^2_{\varepsilon j} + \sigma^2_{\varepsilon,K+1}}.$$

Thus, this holds true for

$$\frac{\left(\sum_{j=1}^K \lambda_j + \lambda_{K+1}\right)^2 \sigma^2_\theta}{\left(\sum_{j=1}^K \lambda_j + \lambda_{K+1}\right)^2 \sigma^2_\theta + \sum_{j=1}^K \sigma^2_{\varepsilon j} + \sigma^2_{\varepsilon,K+1}} < \frac{\left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_\theta}{\left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_\theta + \sum_{j=1}^K \sigma^2_{\varepsilon j}}.$$

Cross-multiplication and canceling common terms yields

$$\left(\sum_{j=1}^K \lambda_j + \lambda_{K+1}\right)^2 \left(\sum_{j=1}^K \sigma^2_{\varepsilon j}\right) < \left(\sum_{j=1}^K \lambda_j\right)^2 \left[\sum_{j=1}^K \sigma^2_{\varepsilon j} + \sigma^2_{\varepsilon,K+1}\right].$$

Expanding the quadratic and canceling common terms yields

$$\left[2\lambda_{K+1} \sum_{j=1}^K \lambda_j + \lambda^2_{K+1}\right] \left(\sum_{j=1}^K \sigma^2_{\varepsilon j}\right) < \left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_{\varepsilon,K+1}.$$

Divide by $\sum_{j=1}^K \sigma^2_{\varepsilon j} > 0$:

$$\lambda^2_{K+1} + 2\lambda_{K+1} \sum_{j=1}^K \lambda_j < \frac{\left(\sum_{j=1}^K \lambda_j\right)^2 \sigma^2_{\varepsilon,K+1}}{\sum_{j=1}^K \sigma^2_{\varepsilon j}}.$$

Solving the quadratic inequality yields

$$\lambda_{K+1} < \lambda_{min} = \left(\sum_{j=1}^K \lambda_j\right) \left[\sqrt{1 + \frac{\sigma^2_{\varepsilon,K+1}}{\sum_{j=1}^K \sigma^2_{\varepsilon j}}} - 1\right].$$

Thus, adding item $K+1$ reduces the reliability of the index when the factor loading λ_{K+1} is low, the sum of the loadings of the other factors $\sum_{j=1}^K \lambda_j$ is high, and the noise of factor $K+1$ is high relative to the noise of the other K factors in the index.

Table A1: Survey Questions: Crosscut Enterprise RCT

Survey Questions: Financial Literacy

Financial Numeracy: Profitability

1) You have the opportunity to buy some young chicks for 100 soles that you can feed with food scraps at no cost, and that will lay eggs bringing you 60 soles each year. You expect the chickens to live two years. What would be your profit?

2) A friend offers to sell you a goat for 400 soles which would give you revenues of 100 soles per year in milk. The goat will be alive for the next 3 years. What would be your profit?

3) The cost of an oven to bake bread is 1,000 soles, and you believe there is a 50% chance it will last one year and a 50% chance it will last two years. The cost of dough is 100 soles per year, but you can sell the baked bread for 900 soles each year. What are your expected profits?

4) You take a loan for 100 soles at 10% simple interest. How much do you owe the bank in one year?

5) You take a loan for 200 soles at 8% simple interest for two years. How much do you owe in two years?

6) You take a loan for 300 soles at 10% simple interest for three years with three equal payments. How much will you pay each year?

7) You have the opportunity to buy a pig for 500 soles certain to live 2 years and provide 400 soles per year. What is the profit?

8) There is an opportunity to buy a sick chicken for 500 soles. There is a 50% chance it will die and a 50% chance it will recover and live two more years providing 400 soles per year. What is the expected profit?

9) A sewing machine costs 1,000 soles and is equally likely to last 4 years or break down after 2 years. You can earn 400 soles per year while it operates. What is the expected profitability?

10) You have no money in your bank account but can purchase a cow for 500 soles that yields 600 soles per year for five years. You consider a loan for 500 soles due in one year at 10% interest. You need 50 soles in your account at all times. (*a*) Which years will you have at least 50 soles? (*b*) How much cash would you have after one year?

11) You have no money in your bank account but can purchase a saw for 2,000 soles for your furniture business that gives returns of 1,500 soles per year. You consider a loan for 2,000 soles due in one year at 10% interest. You need 50 soles in your account at all times. (*a*) Which years will you have at least 50 soles? (*b*) How much cash would you have after one year?

Financial Knowledge

1) How would you rate your current understanding of savings?

2) How would you rate your current understanding of borrowing?

3) How would you rate your current understanding of investing?

4) How often do you engage in financial discussions with others, such as family members or friends?

Financial Behavior

1) Have you ever borrowed money?

2) Do you assess the risk before you make any investment?

3) In the past 12 months have you been personally saving money in any of the following ways? (*a*) Saving cash at home or in your wallet; (*b*) Savings account in a bank or financial institution; (*c*) Giving money to your family to save on your behalf; (*d*) Savings in self-help groups or informal savings clubs.

Financial Competency

1) Who makes day-to-day decisions about how to use money in your household?

2) Do you do any of the following for yourself or your household? (*a*) Make a plan to manage your income and expenses; (*b*) Keep a note of your spending; (*c*) Keep money for bills separate from day-to-day spending money; (*d*) Make a note of upcoming bills so you don't miss them; (*e*) Keep a close personal watch on my financial affairs.

Financial Confidence

1) How confident do you feel in making financial decisions?

2) If you personally faced a major expense today—equivalent to your monthly income—would you be able to pay it without borrowing or asking family or friends to help?

3) If you personally faced a major expense today—equivalent to your monthly income—how confident are you about raising the money if given a day's time?

4) Some people set financial goals, such as paying for kids' education, buying a vehicle, or becoming debt free. Do you (personally, or with your spouse/partner) have any financial goals?

5) I find it more satisfying to spend/invest money than to save it for the long term.

6) I am prepared to risk some of my own money when saving or making an investment.

7) I am satisfied with my present financial situation.

8) My financial situation limits my ability to do the things that are important to me.

9) I have too much debt right now.

Continued on next page...

Table A1 (continued) — Survey Questions: Crosscut Enterprise RCT

Survey Questions: Entrepreneurial Aspirations
Aspirations
Responses on a scale of 0–10 (strongly disagree to strongly agree).
1) All things considered, how happy would you say you are today?
2) All things considered, how hopeful do you feel about the future?
3) It is better to learn to accept the reality of things than to dream for a better future.
4) It is better to have aspirations for your life than to accept each day as it comes.
5) I am satisfied with my current income and spending levels.
6) It is wiser to establish goals than to address situations as they arrive.
7) I have specific goals and plans for my future income.
Agency
How important is each concept to success (0 = not important, 10 = extremely important)?
1) On a scale of 0 to 10, how important is hard work to prosper in making money?
2) On a scale of 0 to 10, how important is being lucky to prosper in making money?
Responses on a scale of 0–10 (strongly disagree to strongly agree).
3) My future is shaped mainly by my own actions rather than by the actions of others.
Pathway Visualization
Responses on a scale of 0–10 (strongly disagree to strongly agree).
1) I can find a way to solve most problems.
2) If my income goes down, I know how to explore new opportunities to bring it up.
3) I become discouraged easily when I encounter obstacles in making money.
4) I know people who I look to as positive examples in how to increase my income level.
5) I understand the different ways and opportunities to increase my income level.

Table A2: Summary Statistics: Three-Country Microenterprise RCT

	Three Countries		India		Peru		Uganda	
	mean	sd	mean	sd	mean	sd	mean	sd
Aspirations Treatment								
Female	0.78	(0.41)	1.00	(0.00)	0.77	(0.42)	0.54	(0.50)
Age	44.22	(11.97)	44.19	(9.11)	47.17	(15.54)	42.04	(11.50)
Education	2.94	(1.58)	2.18	(0.98)	4.83	(1.66)	2.40	(0.71)
Number of Dependents	4.11	(3.43)	2.52	(2.18)	2.69	(1.79)	6.91	(3.71)
Monthly Income (USD)	161.96	(213.49)	80.32	(89.78)	278.61	(337.62)	154.57	(124.30)
Base Financial Literacy	0.11	(0.97)	0.34	(0.90)	-0.04	(1.06)	-0.04	(0.94)
Base Aspirational Hope	0.19	(0.91)	0.33	(0.77)	0.12	(1.20)	0.06	(0.78)
Financial Literacy Treatment								
Female	0.83	(0.38)	1.00	(0.00)	0.81	(0.39)	0.67	(0.47)
Age	46.53	(13.63)	49.69	(9.80)	46.25	(18.30)	43.51	(11.94)
Education	3.01	(1.85)	1.93	(0.98)	5.06	(1.93)	2.46	(0.92)
Number of Dependents	3.44	(3.05)	1.60	(1.33)	2.27	(1.52)	6.13	(3.25)
Monthly Income (USD)	199.83	(274.01)	68.79	(118.58)	350.73	(360.46)	200.33	(234.18)
Base Financial Literacy	-0.05	(1.00)	-0.50	(1.02)	0.36	(0.73)	0.07	(0.99)
Base Aspirational Hope	-0.21	(1.04)	-0.62	(1.19)	0.17	(0.99)	-0.12	(0.75)
Financial Lit × Aspirations								
Female	0.78	(0.41)	1.00	(0.00)	0.52	(0.51)	0.65	(0.48)
Age	46.05	(12.81)	46.01	(10.78)	52.67	(16.27)	40.46	(9.79)
Education	2.86	(1.74)	1.98	(0.95)	5.02	(1.97)	2.48	(0.69)
Number of Dependents	3.04	(2.69)	1.79	(1.53)	2.22	(1.47)	5.81	(2.98)
Monthly Income (USD)	157.03	(173.43)	77.17	(53.11)	254.94	(225.34)	191.94	(187.28)
Base Financial Literacy	-0.20	(0.94)	-0.16	(0.87)	-0.32	(1.03)	-0.14	(0.97)
Base Aspirational Hope	0.03	(0.95)	0.31	(0.98)	-0.22	(1.02)	-0.21	(0.70)
Control								
Female	0.86	(0.35)	1.00	(0.00)	0.74	(0.44)	0.80	(0.41)
Age	47.33	(13.14)	48.74	(8.36)	47.32	(18.74)	45.15	(9.74)
Education	2.98	(1.80)	1.95	(0.78)	4.74	(1.92)	2.27	(0.73)
Number of Dependents	3.04*	(2.66)	1.63	(1.25)	2.66	(1.66)	5.71	(3.34)
Monthly Income (USD)	195.44	(370.49)	59.27	(49.59)	328.50	(533.40)	207.89	(290.53)
Base Financial Literacy	0.11	(1.06)	0.30	(0.98)	-0.11	(1.05)	0.11	(1.14)
Base Aspirational Hope	-0.04	(1.07)	-0.16	(0.70)	-0.15	(1.26)	0.31	(1.21)
Date of Experiment:	-	-	June	2023	July	2023	August	2024
Observations	743		348		126		269	

*significantly different from treatment at the $p < 0.05$ significance level.

Table A3: Impact on Financial Knowledge, Numeracy, Behavior and Digital Literacy

Financial Knowledge	Three Countries	India	Peru	Uganda
Financial Literacy	0.355***	0.318	0.0292	0.605**
	(0.0945)	(0.157)	(0.0678)	(0.108)
Entrepreneurial Aspirations	0.300***	0.269*	0.0484	0.397**
	(0.0746)	(0.0980)	(0.165)	(0.0830)
FinLiteracy x Aspirations	0.351***	0.288	0.374***	0.509**
	(0.0985)	(0.155)	(0.0831)	(0.158)
Financial Numeracy	Three Countries	India	Peru	Uganda
Financial Literacy	0.292**	0.217	0.100	0.588*
	(0.115)	(0.182)	(0.329)	(0.235)
Entrepreneurial Aspirations	-0.147	-0.0863	-0.450*	0.0942
	(0.102)	(0.112)	(0.215)	(0.193)
FinLiteracy x Aspirations	0.209**	0.188	0.319**	0.492*
	(0.0917)	(0.110)	(0.1000)	(0.179)
Financial Knowledge	Three Countries	India	Peru	Uganda
Financial Literacy	0.355***	0.318	0.0292	0.605**
	(0.0945)	(0.157)	(0.0678)	(0.108)
Entrepreneurial Aspirations	0.300***	0.269*	0.0484	0.397**
	(0.0746)	(0.0980)	(0.165)	(0.0830)
FinLiteracy x Aspirations	0.351***	0.288	0.374***	0.509**
	(0.0985)	(0.155)	(0.0831)	(0.158)
Observations	743	348	126	269

Dependent variable formed through the first principal component (PCA).
Multi-country regressions incorporate fixed effects at the country level.
Standard errors clustered at the treatment group level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Impact on Financial Competency, Confidence, and Overall Financial Literacy

Financial Competency	Three Countries	India	Peru	Uganda
Financial Literacy	-0.157**	-0.219***	-0.205	-0.166
	(0.0672)	(0.0350)	(0.258)	(0.149)
Entrepreneurial Aspirations	-0.00521	-0.0285	0.0687	-0.0807
	(0.0727)	(0.0768)	(0.182)	(0.113)
FinLiteracy x Aspirations	-0.0543	-0.00742	-0.362***	0.0879
	(0.0785)	(0.109)	(0.0652)	(0.0976)
Financial Confidence	Three Countries	India	Peru	Uganda
Financial Literacy	-0.0375	0.216**	-0.00198	-0.204*
	(0.0883)	(0.0448)	(0.141)	(0.0833)
Entrepreneurial Aspirations	0.0231	0.203*	-0.0532	-0.0749
	(0.0900)	(0.0665)	(0.219)	(0.0582)
FinLiteracy x Aspirations	-0.00821	0.205**	-0.214	0.0225
	(0.0859)	(0.0509)	(0.163)	(0.0979)
Total Financial Literacy	**Three Countries**	**India**	**Peru**	**Uganda**
Financial Literacy	0.289***	0.258	0.0606	0.349***
	(0.0645)	(0.168)	(0.0615)	(0.0335)
Entrepreneurial Aspirations	0.155*	0.361***	-0.138	0.170
	(0.0792)	(0.0394)	(0.111)	(0.144)
FinLiteracy x Aspirations	0.279***	0.396***	0.0997	0.436*
	(0.0543)	(0.0489)	(0.135)	(0.138)
Observations	743	348	126	269

Dependent variable formed through the first principal component (PCA).
Multi-country regressions incorporate fixed effects at the country level.
Standard errors clustered at the treatment group level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Impact on Entrepreneurial Aspirations, Agency, Visualization of Pathways, and Overall Aspirational Hope: **Factor Analysis-Based Indices**

Entrepreneurial Aspirations	Three Countries	India	Peru	Uganda
Financial Literacy	0.127	0.148	0.454	-0.0780
	(0.127)	(0.124)	(0.257)	(0.268)
Entrepreneurial Aspirations	0.306***	0.231**	0.189	0.332*
	(0.0732)	(0.0612)	(0.265)	(0.128)
FinLiteracy x Aspirations	0.181*	0.0841	0.332	0.105
	(0.0964)	(0.0961)	(0.272)	(0.222)
Entrepreneurial Agency	Three Countries	India	Peru	Uganda
Financial Literacy	0.196	0.432***	-0.117	0.0607
	(0.123)	(0.0487)	(0.137)	(0.309)
Entrepreneurial Aspirations	0.297**	0.388**	0.0378	0.316
	(0.114)	(0.0846)	(0.284)	(0.220)
FinLiteracy x Aspirations	0.218	0.448***	0.160	-0.0989
	(0.135)	(0.0592)	(0.184)	(0.331)
Pathway Visualization	Three Countries	India	Peru	Uganda
Financial Literacy	0.162	0.272	0.564*	-0.144
	(0.142)	(0.124)	(0.244)	(0.257)
Entrepreneurial Aspirations	0.162	-0.0340	0.105	0.292*
	(0.111)	(0.198)	(0.248)	(0.113)
FinLiteracy x Aspirations	0.281**	0.332	0.507**	-0.00877
	(0.126)	(0.185)	(0.177)	(0.213)
Aspirational Hope Index	**Three Countries**	**India**	**Peru**	**Uganda**
Financial Literacy	0.226	0.413**	0.379*	-0.0547
	(0.132)	(0.0775)	(0.197)	(0.323)
Entrepreneurial Aspirations	0.330***	0.235	0.139	0.418**
	(0.0906)	(0.149)	(0.258)	(0.117)
FinLiteracy x Aspirations	0.302**	0.364*	0.406	0.0197
	(0.105)	(0.137)	(0.236)	(0.178)
Observations	743	348	126	269

Dependent variable formed by the first factor using confirmatory factor analysis (CFA).
Multi-country regressions incorporate fixed effects at the country level.
Standard errors clustered at the treatment group level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Impact on Entrepreneurial Aspirations, Agency, Visualization of Pathways, and Overall Aspirational Hope: Robustness Check using **PCA Indices**

Entrepreneurial Aspirations	Three Countries	India	Peru	Uganda
Financial Literacy	0.134	0.173	0.475	-0.0953
	(0.128)	(0.117)	(0.262)	(0.269)
Entrepreneurial Aspirations	0.306***	0.245**	0.188	0.315*
	(0.0713)	(0.0509)	(0.257)	(0.125)
FinLiteracy x Aspirations	0.197*	0.118	0.351	0.0937
	(0.0947)	(0.102)	(0.273)	(0.209)
Entrepreneurial Agency	Three Countries	India	Peru	Uganda
Financial Literacy	0.185	0.430***	-0.138	0.0428
	(0.125)	(0.0523)	(0.144)	(0.308)
Entrepreneurial Aspirations	0.300**	0.408**	0.0228	0.308
	(0.114)	(0.0753)	(0.284)	(0.214)
FinLiteracy x Aspirations	0.216	0.462***	0.144	-0.110
	(0.137)	(0.0532)	(0.173)	(0.328)
Pathway Visualization	Three Countries	India	Peru	Uganda
Financial Literacy	0.177	0.307*	0.567*	-0.148
	(0.144)	(0.122)	(0.249)	(0.260)
Entrepreneurial Aspirations	0.177	-0.0108	0.130	0.291*
	(0.111)	(0.199)	(0.245)	(0.115)
FinLiteracy x Aspirations	0.279*	0.339	0.505**	-0.0255
	(0.132)	(0.201)	(0.175)	(0.213)
Aspirational Hope Index	**Three Countries**	**India**	**Peru**	**Uganda**
Financial Literacy	0.232	0.440***	0.383*	-0.0736
	(0.136)	(0.0681)	(0.199)	(0.328)
Entrepreneurial Aspirations	0.341***	0.261	0.146	0.408**
	(0.0922)	(0.148)	(0.257)	(0.118)
FinLiteracy x Aspirations	0.310**	0.390*	0.408	0.00143
	(0.110)	(0.141)	(0.232)	(0.176)
Observations	743	348	126	269

Dependent variable formed through the first principal component (PCA).
Multi-country regressions incorporate fixed effects at the country level.
Standard errors clustered at the treatment group level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A7: Impact on Entrepreneurial Aspirations, Agency, Visualization of Pathways, and Overall Aspirational Hope: Robustness Check using **Anderson Indices**

Entrepreneurial Aspirations	Three Countries	India	Peru	Uganda
Financial Literacy	0.0876	-0.0955	0.164*	0.263**
	(0.0938)	(0.155)	(0.0832)	(0.0510)
Entrepreneurial Aspirations	0.197**	0.178	0.177	0.237
	(0.0923)	(0.127)	(0.165)	(0.168)
FinLiteracy x Aspirations	-0.0437	-0.212*	0.0813	0.120
	(0.115)	(0.0805)	(0.228)	(0.275)
Entrepreneurial Agency	Three Countries	India	Peru	Uganda
Financial Literacy	0.151	0.469***	-0.276	-0.00846
	(0.132)	(0.0374)	(0.175)	(0.251)
Entrepreneurial Aspirations	0.277**	0.430***	-0.105	0.277
	(0.114)	(0.0459)	(0.297)	(0.160)
FinLiteracy x Aspirations	0.171	0.450***	-0.0874	-0.132
	(0.133)	(0.0526)	(0.174)	(0.256)
Pathway Visualization	Three Countries	India	Peru	Uganda
Financial Literacy	0.202	0.367*	0.566*	-0.152
	(0.147)	(0.117)	(0.257)	(0.263)
Entrepreneurial Aspirations	0.202*	0.0279	0.166	0.290*
	(0.113)	(0.200)	(0.240)	(0.117)
FinLiteracy x Aspirations	0.273*	0.349	0.498**	-0.0573
	(0.143)	(0.228)	(0.174)	(0.213)
Aspirational Hope Index	**Three Countries**	**India**	**Peru**	**Uganda**
Financial Literacy	0.259**	0.552***	0.143	0.00645
	(0.113)	(0.0911)	(0.0867)	(0.171)
Entrepreneurial Aspirations	0.247**	0.206	0.0552	0.356***
	(0.0855)	(0.121)	(0.206)	(0.0582)
FinLiteracy x Aspirations	0.214*	0.323*	0.204	0.0229
	(0.103)	(0.122)	(0.197)	(0.172)
Observations	743	348	126	269

Dependent variable formed using Anderson indices.
Multi-country regressions incorporate fixed effects at the country level.
Standard errors clustered at the treatment group level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$