How robust are composite indicators? An almost certain stochastic simulation approach

Draft. Not for citation

Sam Jones

University of Copenhagen,
Department of Economics

September 2013

Abstract

Stochastic techniques such as Monte Carlo simulations are widely used to investigate complex empirical problems. In particular, they can be used to quantify the sensitivity of composite indicators to changes in underlying parameters (Saisana et al., 2005). One problem with these methods is that the adequacy or completeness of resulting simulated (Monte Carlo) distributions is not clear. Since the concept of robustness refers to the entire space of input parameter choices, this is a critical weakness. To address this gap, two simple non-parametric estimators of the coverage of samples from unknown distributions are proposed. These are combined with stochastic simulation methods to give measures of pointwise dominance that hold to a desired level of certainty. Applications to the Human Development Index and the Alkire-Foster class of multi-dimensional deprivation measures are provided.
1 Introduction

Recent years have seen a flourishing of interest in composite indicators. In part this reflects a desire to make sense of the sheer volume of macro- and micro-economic data that is becoming available. It also reflects a sentiment that trends in complex phenomena are not fully captured by looking at individual variables in isolation. This is not just because some variables may represent poor or noisy proxies for the phenomena of interest. Rather, their joint distribution may be of inherent interest. To take one example, it is generally acknowledged that deprivation is multi-dimensional (Atkinson, 2003). However, the extent to which people are simultaneously deprived in multiple dimensions cannot be gauged from an analysis of the marginal distribution of each dimension of interest.

Composite indicators are not without their critics. Ravallion (2012) raises a range of concerns with ‘mash-up indices of development’, such as a lack of clarity regarding trade-offs between dimensions embodied in any given measure. A further issue is whether conclusions based on one set of parameters used to aggregate across separate dimensions hold over (all reasonable) alternative sets. For instance, if countries are ranked according to some composite indicator, certain country rankings may be highly sensitive to different parameter choices while others may be relatively stable.

The robustness of composite indicators represents the central focus of this study. This interest is not new. Various studies assess the properties of popular composite indexes, as well as the robustness of multidimensional deprivation measures. Following progress in the (consumption) poverty literature, one approach is to develop metrics of distributional dominance based on the multivariate empirical distribution of the underlying data (e.g., Duclos et al., 2006). This is theoretically attractive because dominance results from this type of analysis should apply to a wide range of plausible social welfare functions.

In practice, however, applying measures of distributional dominance to complex multivariate distributions – which might include variables of a binomial, ordinal and continuous nature – remains work in progress. As Arndt et al. (2012a) note, most available approaches place restrictions on the nature of feasible social welfare functions, particularly regarding the sign of cross-derivatives between dimensions. Moreover, despite important advances in extending multivariate measures of stochastic dominance to ordinal variables (see Yalonetzky, 2013), these approaches do not easily transfer to non-smooth functions such as the popular counting-type
measures of multidimensional well-being.¹

Following the above, an alternative route to assessing the robustness of composite indicators is pursued here based on stochastic simulation methods. This does not investigate the distribution of the underlying data directly, but rather considers the sensitivity of the chosen indicator to alternative input parameters. The aim is to characterize the empirical probability density function of the indicator and, thus, ascertain whether inferences based on any particular set of assumptions are fragile. In order to avoid bias from subjective choices of input parameters, stochastic methods such as Monte Carlo simulations or variants thereof are employed to generate a representative sample of the input space (see Section 2).

This broad approach represents a type of sensitivity analysis (for a general overview, including of stochastic techniques, see Saltelli et al., 2009) or, more specifically, an uncertainty analysis.² Uncertainty analysis has particular advantages in situations where closed form analytical tests of a model’s properties are elusive. Not only does it hold the practical attraction of simplicity and transparency; but also, by giving an estimate of the empirical distribution of outputs, it permits exploration of sources of variation or interesting features of this distribution (as might be of interest to a policy-maker). Nonetheless, a fundamental weakness with these methods is they often seem rather perfunctory (Saltelli and Annoni, 2010). The risk is that uncertainty analysis fails to consider a sufficient range of combinations of inputs and, consequently, does not provide an adequate characterisation of the empirical distribution of outputs.

Previous studies have used stochastic uncertainty (sensitivity) analyses to evaluate the properties of composite indicators.³ A leading example is Saisana et al. (2005), who use Monte Carlo techniques to assess the reliability of country rankings on the UN’s technology achievement index. Cherchye et al. (2007) consider a wider range of model uncertainties for the same index. Kovacevic and Aguña (2010) submit the UNDP’s Human Development Index to an uncertainty analysis; and Nussbaumer et al. (2012) investigate the sensitivity of a multi-dimensional energy

¹For instance, Yalonetzky (2011) concludes that: “... traditional dominance conditions ... are not appropriate for poverty counting measures ..., except when extreme poverty identification approaches are considered.” (p. 18).
²The distinction between uncertainty and sensitivity analysis is not strict. Helton et al. (2006) suggests that uncertainty analysis seeks to quantify the degree of uncertainty in outputs due to uncertainty regarding the appropriate values or form of inputs, while sensitivity analysis seeks to determine the sources of variation in outputs due to differences in inputs.
³Robustness of composite indicators to changes in underlying parameters has also been evaluated using other methods, including linear programming. Relevant references are McGillivray and Noorbakhsh (2004); Cherchye et al. (2008); Arndt et al. (2012a); Foster et al. (2013). Space considerations mean that alternative methods cannot be compared. However, an advantage of the approach developed here is that it applies to indicators of any form and arbitrary complexity.
poverty index (MEPI) to alternative dimension-specific weights and cut-off values. Comparing these studies, one notes considerable differences in the number of draws (simulations) employed. For the latter two studies, no explicit justification is provided for the number of draws (5000 and 1000 respectively). In the first two studies, the number of draws is determined with reference to the number of inputs varying on each evaluation. The stated intention is to generate a sample that is both representative and of an adequate size to calculate ancillary variance measures, such as a sensitivity index for each input. For instance, Cherchye et al. (2007) evaluate 24,576 unique input combinations; however, this corresponds to under 1% of over 12.5 million possible combinations (all inputs are discrete).

Regardless of whether the number of simulations is substantiated \textit{ex ante}, none of these (or other) studies refer to \textit{ex post} measures of simulation coverage. This is a relevant omission because, in contrast to general evaluations of the sensitivity of a model’s output to changes in input parameters, strict assessment of robustness requires that conclusions (such as rank order) are consistent across all possible input combinations. Since the nature of the output distribution is unlikely to be known in advance, prior estimates of sample size requirements provide no guarantee that all interesting domains of the parameter space will be covered in a particular random sample. This is of especial relevance where the output distribution is highly skewed, has heavy tails, or is discontinuous over some range.

To address this gap, this paper proposes simple non-parametric methods to evaluate the completeness (coverage) of samples from \textit{ex ante} unknown distributions. It shows how such estimators can be combined with stochastic search techniques to evaluate the robustness of composite indicators of any form. In turn, the study shows that by simulating the empirical distribution of outputs to an arbitrary estimated degree of completeness, ‘almost certain’ pointwise dominance comparisons can be made. Bilateral comparisons of this type are easy to communicate and have direct implications for stochastic dominance at all orders (see Hansen et al., 1978). Together this provides a general method that responds to Ravallion’s 2012 call for better use of technology to verify the robustness of composite indicators.

The remainder of the paper is structured as follows: Section 2 gives a more formal statement of the ‘robustness problem’ and discusses the strengths and weaknesses of undertaking a stochastic search of the input parameter space. Section 3 presents two non-parametric estimators of sample coverage. To guide practitioners, Section 4 combines these estimators with stochastic search to give a general algorithm for simulating a composite indicator to within a desired level of certainty (completeness). Section 5 applies these methods to two well-known examples: (i) the
UNDP’s Human Development Index (HDI); and (ii) the Alkire-Foster class of multi-dimensional deprivation measures, employed specifically to analyse trends in well-being in Mozambique. Section 6 concludes.

Before proceeding, a few additional comments are in order. First, the proposed estimators of sample coverage are not entirely original. Nonetheless, they are not widely known beyond specific areas of application and, to the author’s knowledge, previously have not been adapted to stochastic simulation contexts. Second, there are many sources of uncertainty in the construction of composite indicators. These include measurement error, choices about how variables are constructed, and weighting schemes. Whilst the present focus is on the parameters of the composite indicator function, stochastic methods easily extend to other sources of uncertainty (e.g., see Saisana et al., 2005).

Third, as already intimated, existing scholarship considers the robustness properties of the two specific indicators reviewed in Section 5. Adequate reference to this literature is beyond the scope of the present exercise and therefore is omitted. Even so, it bears remarking there is no generally established methodology for assessing the robustness of either indicator and that methods employed for one do not necessarily transfer to the other. Additionally, existing approaches can be difficult to implement and/or communicate to non-specialist audiences, limiting their broader appeal. This provides a further motivation for applying stochastic simulation techniques.

Finally, the estimators proposed here are not restricted to evaluating composite indicators. Ex post measures of sample coverage may be informative in a wide range of situations where stochastic simulations can be employed. For instance, Edward Leamer has advocated use of sensitivity analysis to assess the fragility of regression models (see Leamer, 1985; Levine and Renelt, 1992). However, as demonstrated by Sala-i-Martin (1997), exhaustive testing of all feasible model specifications can be extremely costly. Thus, stochastic search combined with estimates of distributional completeness (e.g., of the beta coefficients on a fixed variable of interest) may be helpful in econometric robustness testing. Further applications and extensions remain for future research.

2 Preliminaries

The ‘robustness problem’ for composite indicators can be clarified via the following definitions:

**Definition 1.** A composite indicator is a mapping \(W : (\theta, X_i) \rightarrow y_i \subset \mathbb{R}\), where:
• \( \theta \in \mathcal{S} = \{ \Theta_1, \Theta_2, \ldots, \Theta_k \} \subset \mathbb{R}^k \) is a draw from the set (S) of feasible input parameters which is assumed to be compact, such that \( \forall \ k : \Theta_k = [a_k, b_k] \), and takes an associated probability distribution \( \Omega \). Constraints on the parameters, such as adding-up requirements, are implicit in the definition of \( \mathcal{S} \).

• \( X_i \) is the raw data for unit \( i \), representing the underlying marginal indicators that are to be aggregated. Throughout, this raw data is taken as given.

• \( y_i \) is assumed to follow a stable distribution, implying \( W \) is well-behaved.

**Definition 2.** Random variable \( y_i \) is said to pointwise dominate (PD) random variable \( y_j \) if \( y_i(\theta; X_i) \geq y_j(\theta; X_j) \) for all feasible values of \( \theta \) in the parameter space and \( y_i(\theta; X_i) > y_j(\theta; X_j) \) for at least one draw of \( \theta \) (for further discussion of PD see Hansen et al., 1978).

**Definition 3.** For a given composite indicator defined by \( W \), a comparison between two distinct units is said to be strictly robust if, \( \forall \ \theta \in \mathcal{S} : (y_i \succ_p y_j) \lor (y_j \succ_p y_i) \), where \( \succ_p \) denotes pointwise dominance.

**Remark.** A simple pointwise measure of the extent or degree of robustness of a comparison between composite indicators can be calculated as:

\[
    r_{ij} = \int_{\theta \in \mathcal{S}} I[y_i(\theta; X_i) - y_j(\theta; X_j) \geq 0] \, d\theta
\]

which reports the share of points in the \( k \)-dimensional parameter space over which \( i \) weakly dominates \( j \). Evidently, if \( (r_{ij} = 1) \cap (r_{ji} \neq 1) \) then \( i \) strictly dominates \( j \) pointwise; however, if \( (0 < r_{ij} < 1) \cap (0 < r_{ji} < 1) \) then a conclusion of non-dominance is warranted.

The above definitions indicate that, for a given composite indicator, if the corresponding input parameter space is discrete and computably enumerable, then the robustness of comparisons between two or more units can be calculated using all possible combinations of the input parameters. This would be equivalent to a brute force solution to a combinatorial search problem. Of course, this situation rarely holds in practice. In most cases input parameters refer to dimension-specific weights, which can take an infinite number of values. Also, even if some (e.g., precision) constraints are imposed that ensure the parameter space is discrete, typically it will not be practical to enumerate and evaluate all feasible input combinations.

Assuming the parameter space cannot be explored exhaustively, other strategies must be employed. As recognized in diverse fields, stochastic methods represent a simple yet powerful
means to explore unknown spaces (Spall, 2003). Indeed, provided there is a non-zero probability of visiting every point in the search space, even pure random search guarantees that the probability of visiting a specified neighbourhood of a unique point (e.g., \( y^* \pm \epsilon \)) converges to one as the number of draws increases (Solis and Wets, 1981; Zabinsky, 2011). Stochastic methods are particularly well suited to address ill-structured, high-dimensional problems; and no specific assumptions need be made about the properties of the function to be evaluated (see James, 1980). These established results mean that for a given prior distribution over input parameters (\( \Omega \)), a stochastic (e.g., Monte Carlo) simulation of the outcome distribution will converge to the ‘true’ distribution with probability one.

A principal disadvantage of stochastic search is that the extent of convergence is difficult to assess. That is, although stochastic search methods have desirable limit properties, a practical problem is to identify when the simulated outcome distribution provides adequate coverage of the full search space, making further search unnecessary. This concern reflects the stopping problem in (stochastic non-linear) optimization. While this is an active research area, stopping rules developed in these contexts typically focus on estimates of the error of a single optimum (the optimality gap), not the completeness of the overall simulated distribution (e.g., Hutchison and Spall, 2005). Also, since the true optimum is by definition unknown, such stopping rules typically rely on asymptotic properties that may be inappropriate in finite samples. Consequently, the task here is to find a reliable heuristic to terminate the search process, the objective being to stop when the expected rate of improvement in our knowledge of the overall simulated distribution falls below a pre-specified threshold. The next section develops such a general rule.

3 Estimating simulation coverage

3.1 Missing mass

To begin, consider a version of the classic balls-in-boxes problem in which a sample of \( N \) balls is thrown independently at a finite series of boxes with probability \( p_j \) of hitting the \( j^{th} \) box. Assuming the total number of distinct boxes is unknown, the problem is to estimate the number of (unoccupied) boxes based on the number that have been occupied, which can be counted. This ‘missing mass’ problem arises in numerous applications such as counting the number of animal species, numismatics, estimating words in a language corpus, and in database query optimization (e.g., Efron and Thisted, 1976; Chaudhuri et al., 1998).
More formally, let $S$ be a discrete, countable set and $s_1, s_2, \ldots, s_N$ independent random draws from $S$ according to the probability measure $\Omega$. Thus, the total probability mass of the elements of $S$ not observed after $N$ draws (missing mass) is given by: $U = P(S \setminus \{s_1, s_2, \ldots, s_N\})$. Following Berend and Kontorovich (2012), it is helpful to note that the expected value of $U$ converges to zero as $N \to \infty$. This can be seen from:

$$E(U) = \sum_{s \in S} p_s (1 - p_s)^N$$

(2)

where $p_y = P(\{s\})$ and the final term is simply the probability mass of a binomial variable seen zero times in $N$ trials. This confirms that stochastic search of $S$ will provide complete coverage of all points in the limit.

Various estimators for $U$ have been proposed (for surveys see Bunge and Fitzpatrick, 1993; Gandolfi and Sastri, 2004). The most famous and widely used take the following form:

$$\hat{T} = \frac{N}{N - c_1} \left( C + c_1 \gamma^2 \right)$$

$$\hat{U} = 1 - C/\hat{T}$$

(3)

(4)

where $\hat{T}$ is an estimate of the total number of boxes (occupied or not), $C$ is the number of boxes occupied after $N$ throws, $c_1$ is the number of boxes occupied by only one ball and $\gamma^2$ is a parameter, to be estimated, that captures variation in the size of the boxes (their occupancy probabilities, $p_s$).

Setting $\gamma^2 = 0$, which is consistent with the assumption that $\Omega$ is a uniform distribution (all boxes are equally likely to be occupied), yields the Good-Turing estimator developed by I. J. Good and Alan Turing as part of their work to crack ciphers for the Enigma machine during World War II (see Good, 1953; Gale and Sampson, 1995). Under this assumption, equation (4) simplifies to $\hat{U} = c_1/N$, which can be interpreted as the empirical likelihood of observing a new (previously unseen) box on the next throw.

Existing studies investigate the properties of the Good-Turing estimator. These show that although it is biased in small samples, it has strong consistency and good convergence properties (McAllester and Schapire, 2000; Wagner et al., 2006). Under relatively weak assumptions, it also is suitable to estimate the probability of rare events (Wagner et al., 2011). These properties make it attractive in the present case where unusual (but important) outcomes may only occur in specific sub-spaces of $S$. 

7
Alternative assumptions for $\gamma^2 > 0$ have been suggested to address some of the shortcomings of the simple Good-Turing estimator. As evident from equation (3), these necessarily generate more conservative missing mass estimates; this is because they allow the probability of occupying the unobserved boxes to be lower than the probability of occupying observed boxes. Defining $\gamma^2$ as the estimated coefficient of variation of the box sizes (i.e., the number of balls in each occupied box) yields one estimator due to Chao and Lee (1992). More recently, Gandolfi and Sastri (2004) adopt a Bayesian approach and propose an estimate for $\gamma^2$ that is bounded between zero and one.\footnote{Specifically: $\gamma^2 = \left( -N \cdot c_1 - C \cdot c_1 + c_1^2 + c_1 \cdot \sqrt{5 \cdot N^2 + 2N(C - 3 \cdot c_1) + (C - c_1)^2} \right) / (2 \cdot N \cdot c_1)$} Simulation results reported by the same authors indicate this estimator performs best when the underlying population is non-uniform.

Application of missing mass estimates to stochastic simulations is straightforward in the case of discrete (feasibly countable) outcomes. For outcomes that are continuous (uncountable), the number of boxes is infinite and therefore missing mass estimates will not approach zero as the number of draws increases. Nonetheless, in most empirical applications the outcome space can be appropriately discretized without loss of meaningful information. This reflects the point that empirical outcomes are hardly ever either estimated or meaningful at an infinite degree of accuracy (for discussion see Bedeian et al., 2009). Thus, simulated outcome values cannot be considered statistically distinguishable beyond a certain level of precision. Typically, the magnitude of measurement uncertainty can be ascertained directly from estimated standard errors on the individual outcome values. In other words, output values can be treated as equivalent when they fall within $z > 0$ standard errors ($\hat{\sigma}$) of a given point, yielding a rounding rule of: $b = -\text{floor}(\log_{10} z \hat{\sigma})$ significant digits.

To avoid arbitrary choices for $z$, guidance can be taken from Sheppard’s approximate correction for calculating distributional moments based on a discrete approximation to a continuous random variable (see Wilrich, 2005). The simple correction for the second moment is given by:

$$\hat{\sigma}^2 \approx \sigma_{\text{discrete}}^2 - \mu^2 / 12$$

where $\mu \equiv 10^b$ is the interval (bin) width used to discretize the continuous variable. Thus, setting the maximum bias to $\epsilon > (\hat{\sigma}_{\text{discrete}}^2 / \sigma^2 - 1)$, the corresponding value for $b$ can be derived as:

$$b > -\log_{10} \left( \hat{\sigma} \sqrt{12\epsilon} \right)$$

(5)

So, for $\epsilon = .02$, a conservative binning interval of $z \approx 1/2$ standard errors is recommended.
To summarise, estimates of missing distributional mass provide a direct indication of the
distributional completeness of a given Monte Carlo sample; thus, they provide a valuable basis
for terminating stochastic searches. Concretely, when the estimated share of unseen points
in the sample space falls below an $\alpha$ threshold, resulting calculations from the distribution of
actually-observed points can be asserted with at least $1 - \alpha$ confidence. This approach strictly
holds in (large) discrete search spaces. However, once measurement uncertainty is considered –
which is almost definitely material in empirical applications of composite indicators – continuous
outcomes can be treated as discrete after appropriate rounding methods are applied.

### 3.2 Distributional distance

An alternative means to assess the completeness of a Monte Carlo simulation is a metric of
distance between a given outcome distribution observed after consecutive blocks of iterations.
Under the assumption that the simulated outcome distribution is stable (see Section 2), then as
the number of iterations increases, the distance between any given quantile of the simulated
distribution at iteration $n$ and the same quantile at iteration $n+k$ will converge to zero. Intuitively,
this represents a generalization of the convergence properties of order statistics employed to
guide stochastic optimization methods (see Zhigljavsky and Hamilton, 2010).

Defining the quantile function of distribution $y$, with CDF $F_y$, evaluated at percentile $p$, as
$Q(y; p) = \inf \{ x \in \mathbb{R} : p \leq F_y(x) \}$, a normalized metric of empirical distance between distri-
bution $z$ and a reference distribution $y$, is:

$$D(z, y; \alpha) = \left[ \int_0^1 \left| \frac{Q(z; p) - Q(y; p)}{Q(y; p)} \right|^\alpha dp \right]^{1/\alpha} \tag{6}$$

This estimator is a modification of the approach proposed in a slightly different context by
Chun et al. (2000). The numerator gives the Minkowski class of distances. Euclidean distance
corresponds to $\alpha = 2$; Chebyshev distance corresponds to $\alpha = \infty$, for which the estimator
becomes: $\sup_{p \in P} |Q(z; p)/Q(y; p) - 1|$. Empirically, these measures can be approximated in
discrete form using a grid of percentiles. However, no discrete approximation to the continuous
outcome distribution is required; thus, this estimator represents a valuable robustness-check on
the missing mass approach.
4 Generic simulation algorithm

Before applying these methods, it helps to combine these estimators with stochastic simulation techniques to give a general algorithm that simulates the empirical distribution of a chosen composite indicator \((y)\). Using previous notation, and denoting \(G\) as the set of units (e.g., countries, regions) over which the composite indicator will be calculated, the pseudo-code runs as follows:

**Algorithm 1** Simulate the distribution(s) \(\forall g \in G : y_g = f(\theta; X_g)\)

1: \{**Preliminaries**\}
2: \textbf{set} iteration block length := \(k\)
3: \textbf{set} maximum number of blocks := \(n\)
4: \textbf{set} distance class := \(\alpha\)
5: \textbf{set} rounding rule := \(b\)
6: \textbf{set} distributional distance tolerance := \(0 \leq \lambda_1 \leq 1\)
7: \textbf{set} missing mass tolerance := \(0 \leq \lambda_2 \leq 1\)
8: \textbf{draw} an \(nk \times m\) matrix of inputs := \(\Theta\)
9: \textbf{initialize} a null \(nk \times g\) output matrix := \(Y\)
10: \{**Iterations**\}
11: \textbf{set} iteration index: \(i \leftarrow 0\)
12: \textbf{for} \(v = 1 : n\) \textbf{do}
13: \hphantom{11:} \textbf{for} \(w = 1 : k\) \textbf{do}
14: \hphantom{11:13:} \(i \leftarrow i + 1\)
15: \hphantom{11:13:} \textbf{calculate} \(\forall g \in G : Y_{ig} \leftarrow f(\theta_i; X_g)\)
16: \hphantom{11:} \textbf{end for}
17: \hphantom{11:} \textbf{calculate} missing mass: \(U_{vg} \leftarrow m(\{Y_{(1:i),g}\}; b)\)
18: \hphantom{11:} \textbf{calculate} distributional distances: \(D_{vg} \leftarrow d(\{Y_{(1:w \times v - 1),g}\}; \{Y_{(1:i),g}\}; \alpha)\)
19: \hphantom{11:} \textbf{if} \(\forall g \in G : U_{vg} < \lambda_1 \text{ and } D_{vg} < \lambda_2\) \textbf{then exit}
20: \textbf{end for}
21: \{**Finish**\}
22: \textbf{return} \(Y_{(1:i),g}\)

Three aspects of this algorithm merit comment. First, for each draw \(\theta_i\), the same input parameters are used to calculate the composite indicator across the \(G\) units. This precisely enables pointwise comparisons to be made between groups – i.e., separate points of the sample space correspond to rows of the output matrix \(Y\). Second, the algorithm does not specify how the matrix of inputs is drawn. Whilst various approaches can be employed (see Helton et al., 2006), pseudo-random sequences are often preferred as they tend to ensure more systematic coverage of multi-dimensional spaces compared to basic random number generators (see Jäckel, 2002; Bhat, 2003). Examples include (shuffled) Halton sequences or Latin Hypercube Sampling, both of
which are generally available in standard software packages. Third, in certain applications users may simply wish to perform a fixed number of total iterations, which is equivalent to setting the tolerance thresholds to zero. Either way, estimates of missing mass or distributional distance calculated after the final iteration correspond to *ex post* measures of the completeness of the simulated distribution(s).

5 Applications

5.1 Human Development Index

Launched in 1990, the UNDP’s Human Development Index (HDI) is a well-known composite indicator of aggregate development progress. The current methodology, updated in 2010, calculates the index as the geometric mean of three sub-indices covering life expectancy \( l \), education \( e \) and income \( y \). Whilst the education index is itself a composite index of two indicators – mean years of schooling and expected years of schooling – for simplicity here each of the three ‘final’ indices is taken as given.\(^5\) On this basis, the official HDI \( h^* \) for country \( i \) is:

\[
h^*_i = (l_i \cdot e_i \cdot y_i)^{1/3}
\]

which is equal to the inverse logarithm of an equal weighted sum of the logarithms of the three sub-components. Thus, a more general form for the index is given by:

\[
h_i = \exp[\omega_1 \ln(l_i) + \omega_2 \ln(e_i) + \omega_3 \ln(y_i)], \quad \sum_{j=1}^{3} \omega_j = 1, \quad w_j > 0 \quad \forall \ j
\]

Focussing on the most recent HDI data from 2012 (UNDP, 2013), the aim is to systematically examine the robustness of country rankings to alternative weights. Following previous sections, I implement a stochastic simulation and associated measures of coverage. This involves three main steps: (i) using a shuffled Halton sequence, I create a \( 5000 \times 3 \) matrix of pseudo-random draws on the unit simplex which, for each row, are subsequently re-scaled to sum to one; (ii) for each weight vector, I use equation (8) to calculate the raw \( h_i \) and associated rank for each of the 187 countries in the dataset; and (iii) with the resulting country-specific raw outcome vectors

\(^5\)In doing so, I also ignore implicit weights associated with how these indices are constructed (Kovacevic and Aguña, 2010), as well as other sources of uncertainty such as due to measurement error (Wolff et al., 2011).
(the simulated distributions), I calculate individual measures of missing mass and distributional distance.

The choice of 5000 iterations may appear ad hoc. *Ex ante* this figure appears plausible, but there are no particular computational or theoretical reasons to exclude higher or lower numbers. The key issue is whether the simulated outcome distributions are *ex post* adequate in the sense of providing a desired level of coverage of the outcome space. This can be verified by recourse to the estimators described in Section 3. To implement the missing mass estimator, the continuous outcome distribution must be treated as-if discrete. To do so, I rely on guidance provided by UNDP which states that HDI values cannot be meaningfully compared at a precision of more than three digits.\(^6\) This choice is supported by the data. Crude estimates of equation (5), using \(\epsilon = 0.02\) and based on the standard errors (by country) from the simulations themselves, suggests a rounding rule of at most three decimal places in all cases.

The resulting estimates of simulation coverage are shown in Table 1 for three countries, calculated at specific iteration points. Cameroon and Nepal are chosen because they respectively display the smallest and largest values for the Good-Turing measure of missing mass after 5000 iterations.\(^7\) Indonesia is chosen as it provides estimates for the same measure that is closest to the average across all 187 countries. Taken together, these countries provide a good indication of the overall performance of the simulation.

Table 1 indicates that the estimated missing distributional mass \((U_{GS})\) is less than 10% after 5000 iterations for all countries; and, for the majority of countries, the share is less than 5%. This implies that pointwise comparisons (as per equation 1) can be asserted with at least 90% confidence. The distributional distance measures \((D_0)\), although not directly comparable in magnitude, tell a similar story to the missing mass estimates. Consistent with the \(U_{GS}\) estimates, Nepal shows the largest normalized distance measures throughout. These show that on comparing Nepal’s simulated distribution after successive blocks of iterations, the maximum proportional change in estimates of any individual quantile equals 2.88% after 1000 iterations, falling to 0.06% after 5000 iterations. Thus, according to both coverage measures, proceeding to more than 5000 draws is unlikely to contribute substantive additional knowledge concerning the distribution of HDI values for each country. That is, the actual draws provide adequate coverage of the outcome space.


\(^7\)Throughout I use the modified Good-Turing estimator due to Gandolfi and Sastri (2004). Estimates for the unmodified Good-Turing estimator (setting \(\gamma^2 = 0\)) are smaller in all cases.
<table>
<thead>
<tr>
<th>Iteration</th>
<th>Cameroon</th>
<th></th>
<th></th>
<th>Nepal</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$U_{GS}$</td>
<td>$D_0$</td>
<td>$U_{GS}$</td>
<td>$D_0$</td>
<td>$U_{GS}$</td>
<td>$D_0$</td>
</tr>
<tr>
<td>1</td>
<td>100.0</td>
<td>–</td>
<td>100.0</td>
<td>–</td>
<td>100.0</td>
<td>–</td>
</tr>
<tr>
<td>500</td>
<td>8.8</td>
<td>9.18</td>
<td>17.1</td>
<td>15.92</td>
<td>28.8</td>
<td>39.11</td>
</tr>
<tr>
<td>1000</td>
<td>9.1</td>
<td>0.64</td>
<td>12.7</td>
<td>1.40</td>
<td>18.0</td>
<td>2.88</td>
</tr>
<tr>
<td>1500</td>
<td>4.9</td>
<td>0.05</td>
<td>9.2</td>
<td>0.17</td>
<td>13.2</td>
<td>0.82</td>
</tr>
<tr>
<td>2000</td>
<td>4.7</td>
<td>0.22</td>
<td>7.5</td>
<td>0.15</td>
<td>11.0</td>
<td>0.54</td>
</tr>
<tr>
<td>2500</td>
<td>3.2</td>
<td>0.04</td>
<td>7.4</td>
<td>0.60</td>
<td>10.3</td>
<td>1.47</td>
</tr>
<tr>
<td>3000</td>
<td>2.4</td>
<td>0.08</td>
<td>7.2</td>
<td>0.05</td>
<td>10.1</td>
<td>0.40</td>
</tr>
<tr>
<td>3500</td>
<td>0.8</td>
<td>0.03</td>
<td>6.5</td>
<td>0.05</td>
<td>8.9</td>
<td>0.41</td>
</tr>
<tr>
<td>4000</td>
<td>0.8</td>
<td>0.02</td>
<td>5.2</td>
<td>0.05</td>
<td>8.8</td>
<td>0.06</td>
</tr>
<tr>
<td>4500</td>
<td>0.8</td>
<td>0.04</td>
<td>4.6</td>
<td>0.04</td>
<td>9.1</td>
<td>0.10</td>
</tr>
<tr>
<td>5000</td>
<td>0.8</td>
<td>0.02</td>
<td>4.3</td>
<td>0.03</td>
<td>8.6</td>
<td>0.06</td>
</tr>
</tbody>
</table>

HDI ($h^*$) 49.5 62.9 46.3
Sim. range 7.9 22.3 39.5

Source: own calculations.

Notes: $U_{GS}$ is the modified Good-Turing estimator due to Gandolfi and Sastri (2004), calculated after the number of iterations indicated in each row; $D$ is a measure of distributional distance calculated according to a discrete version of equation (6), where the reference distribution is the outcome vector observed after the number of iterations indicated in the preceding row; $D$ also uses a grid of 50 percentile points and sets $\alpha = 0$. 

Table 1: Estimates of sample coverage for HDI simulations
To get a better sense of these results, Figure 1 provides a visual comparison of the coverage of the simulated distributions after 50, 500, and 5000 iterations for Cameroon and Nepal. In the case of Cameroon (panel a) it shows the range of simulated outcomes is symmetrical and relatively tightly packed about the mean. Thus, even after 500 iterations the estimated density appears relatively smooth and, with the exception of somewhat more complete coverage as indicated by the stripchart, additional iterations provide little change in our overall appreciation of the distribution. For example, after 500 iterations the estimated range of the HDI for Cameroon is 7.3, which increases marginally to 7.9 after 5000 iterations. The case of Nepal is quite different. The figure (panel b) indicates a long right tail and a much larger range of values taken by the index. It follows that the simulation shows quite patchy coverage at 500 iterations; and even after 5000 iterations the corresponding stripchart indicates various gaps. These results are highly consistent with the estimates of missing mass. Nonetheless, it is worth highlighting that the range of values taken by the HDI for Nepal is well approximated after 500 iterations at 36.4, compared to 39.6 after 5000 iterations. This explains the greater similarity in the distributional distance results across the three countries on completion of the simulation (Table 1).

Turning to the robustness of country rankings, for illustrative purposes I focus on a select
Figure 2: Ranges of simulated HDI rankings (5000 draws), selected countries

Source: own calculations.
Table 2: Summary of bilateral pointwise rank comparisons, selected countries

<table>
<thead>
<tr>
<th></th>
<th>Norway</th>
<th>Qatar</th>
<th>Panama</th>
<th>Peru</th>
<th>China</th>
<th>Botsw.</th>
<th>India</th>
<th>Nigeria</th>
<th>Liberia</th>
<th>Niger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>.</td>
<td>4</td>
<td>-37</td>
<td>-62</td>
<td>-75</td>
<td>-61</td>
<td>-124</td>
<td>-143</td>
<td>-149</td>
<td>-156</td>
</tr>
<tr>
<td>Qatar</td>
<td>94.26</td>
<td>.</td>
<td>47</td>
<td>34</td>
<td>-3</td>
<td>15</td>
<td>-34</td>
<td>-36</td>
<td>-44</td>
<td>-76</td>
</tr>
<tr>
<td>Panama</td>
<td>100</td>
<td>80.06</td>
<td>.</td>
<td>-10</td>
<td>-31</td>
<td>-3</td>
<td>-74</td>
<td>-82</td>
<td>-90</td>
<td>-119</td>
</tr>
<tr>
<td>Peru</td>
<td>100</td>
<td>91.92</td>
<td>100</td>
<td>.</td>
<td>-7</td>
<td>19</td>
<td>-51</td>
<td>-62</td>
<td>-78</td>
<td>-93</td>
</tr>
<tr>
<td>China</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>.</td>
<td>32</td>
<td>-27</td>
<td>-32</td>
<td>-41</td>
<td>-72</td>
</tr>
<tr>
<td>Botsw.</td>
<td>100</td>
<td>99.32</td>
<td>100</td>
<td>96.06</td>
<td>79.90</td>
<td>.</td>
<td>35</td>
<td>-1</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>India</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>86.74</td>
<td>.</td>
<td>-1</td>
<td>-10</td>
<td>-32</td>
</tr>
<tr>
<td>Nigeria</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>.</td>
<td>14</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Liberia</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.52</td>
<td>100</td>
<td>93.4</td>
<td>.</td>
<td>6</td>
</tr>
<tr>
<td>Niger</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.96</td>
<td>100</td>
<td>99.72</td>
<td>91.12</td>
<td>.</td>
</tr>
</tbody>
</table>

Source: own calculations.

Notes: figures above the diagonal indicate the most favourable rank comparison of the column country against the row country, where positive values imply the column country has a superior rank to the row country over all sample points; figures below the diagonal report the share of weight vectors over which the rank of the column country is better than (lower) or equal to the rank of the row country.

Figure 2 plots the range of HDI rankings taken by ten select countries (spanning low, middle and high income categories) over all 10000 simulations. Two points merit comment. First, the ranges of the country-specific rankings are not uniform, indicating considerable differences between countries in the sensitivity of their ranks to alternate weighting choices. This is driven by the extent to which each country takes high (low) values across all sub-indices or, as in the cases of Botswana and Qatar, they perform well on some sub-indices but poorly on others. Naturally, results for these latter countries will be more sensitive to alternative weight vectors. Second, countries with proximate ranks according to the base weight vector ($h^*$) show substantial overlap in the range of rankings over the simulations. However, this does not apply generally – for instance, the worst (numerically highest) rank taken by Norway is superior (lower) to the best ranking of all other selected countries, excluding Qatar. This means that Norway’s HDI is superior to the HDI of other countries, even when different weight vectors are applied.

Figure 2 provides a crude guide to bilateral dominance conditions. This is because it does not reflect pointwise comparisons. Table 2 summarises results for pointwise rank comparisons for each bilateral pair. The top half of the table (above the diagonal) reports the most favourable rank comparison of the column country against the row country. The reported value is the number of positions by which the column country is ranked better than the row country for

8Full details available from the author on request.
the most favourable of all weight vectors (applying the same weights to both countries). More precisely, it gives the minimum of the row country rank minus the column country rank over all simulated points. Negative values imply that for no weight vector does the column country have a superior (numerically lower) rank than the row country. In turn, the bottom half of the table reports results according to equation (1). It gives the share of points (weight vectors) for which the rank of the column country is better than (lower) or equal to the rank of the row country.

These results are considerably more nuanced those implied by Figure 2, underlining the merit of pointwise comparisons. For example, whilst the range of rankings taken by Nigeria approximately appear to be a subset of the rankings taken by Botswana, the pointwise comparison indicates there is no weight vector for which Nigeria’s HDI rank is superior to that of Botswana. In fact, the closest position is such that Botswana is ranked one position higher than Nigeria. Two further comparisons are instructive. First, despite the wide range of ranks taken by Qatar, it dominates all other countries except Panama and Norway for at least 90% of weight vectors. Second, although Niger (ranked bottom according to $h^*$) achieves a superior ranking to Botswana and Nigeria on at least one weight vector, the probability that Niger ranks superior to either of these countries on a randomly chosen weight vector is less than 1%.

The proportions reported in Table 2 are meaningful per se. Even so, estimates of their statistical significance can be calculated using a simple counting approach (e.g., Wilks, 2006). To investigate whether $y_i$ pointwise dominates $y_j$, a reasonable ‘non-dominance’ null hypothesis is that the observed share of such dominant points is less than or equal to $\pi \leq 1$. Assuming each sampled point is independent, the expected count of dominant points follows a Binomial distribution with parameters equal to the sample size and $\pi$. For example, Nigeria’s HDI pointwise dominates that of Niger over 4986 sampled points (99.72% of trials). Assuming a null hypothesis that the share of dominant points is less or equal to 99.5%, a one-sided exact binomial test gives a probability of 1.2%. Thus, at significance levels above 1%, the null can be rejected in favour of the alternative hypothesis – i.e., that the true share of dominant points is at least 99.5%. This procedure is useful to resolve border-line dominance cases. However, more broadly, it indicates a range of analytical procedures can be applied to evaluate pointwise dominance.

\[9\text{ Implemented in the R environment with the command: } \text{binom.test}(4986,5000,.995,\text{alternative="greater"}).\]
5.2 Alkire-Foster class of multidimensional poverty measures

A popular microeconomic composite indicator is the Alkire-Foster (AF) class of deprivation measures (Alkire and Foster, 2011), which represents a generalization of the Foster-Greer-Thorbecke class (Foster et al., 1984). While the latter is typically used to consider consumption poverty, the former applies over multiple domains of deprivation.

The AF class of measures can be stated in general form via four definitions:

\[
d_{ij} = \begin{cases} 
0 & \text{if } x_{ij} > \bar{z}_j \\
1 & \text{otherwise}
\end{cases} \quad (9)
\]

\[
g_{ij} = \begin{cases} 
0 & \text{if } d_{ij} = 0 \\
(\bar{z}_j - x_{ij})/\bar{z}_j & \text{otherwise}
\end{cases} \quad (10)
\]

\[
h_i = \mathcal{I}\left(\sum_j \omega_j d_{ij} > \kappa \right) \quad (11)
\]

\[
M_{\alpha,\beta,I} = \frac{1}{|I|} \sum_{i \in I} h_i \left[ \sum_j \omega_j d_{ij} (g_{ij})^{\alpha} \right]^{\beta} \quad (12)
\]

\[\sum_j \omega_j = 1, \; \omega_j > 0 \; \forall \; j, \; 0 < \kappa \leq 1, \; \alpha \in \{0, 1, 2\}, \; \beta \in \{0, 1\}\]

Equation (9) indicates whether unit \(i\) is deprived with respect to domain \(j\), taking a value of one if unit \(i\) is observed below the domain-specific deprivation threshold \((\bar{z}_j)\). In principle, these thresholds could be treated as a source of uncertainty and, thus, permitted to vary in stochastic simulations. However, the question of how deprivation is defined in individual domains is a generic issue that is not particular to the multidimensional setting.\(^{10}\) Also, in empirical applications such as here, where binary indicators are employed (e.g., access to clean water), thresholds cannot be meaningfully varied. As a result, and to ensure simplicity, the thresholds are held fixed.

Equation (10) is the (non-negative) deprivation gap expressed as a percentage of the deprivation threshold.

\(^{10}\)As Alkire and Foster (2011) put it, “... [for] many variables there is a general understanding of what a given cutoff level means and how to go about selecting it.” (p. 482).
threshold; and equation (11) is a binary variable taking the value of one if the weighted sum of deprivation indicators falls above a specified threshold, \( \kappa \). Identification of units as deprived on individual domains, followed by identification of their multidimensional welfare status based on a linear combination of these domains reflects the dual cut-off approach that is a core feature of the method. Finally, equation (12) aggregates these variables over all units in a given group \((I)\). The particular form of aggregation depends on choices for parameters \( \alpha \) and \( \beta \). Setting \( \beta = 0 \) fixes the term in square brackets to one, which reduces the measure to a simple headcount (the share of units defined as deprived in multiple dimensions). Setting \( \beta = 1, \alpha = 0 \) gives the adjusted headcount; larger values for \( \alpha \) correspond to adjusted poverty gap measures.

It is evident from these expressions that, aside from specific choices for \( \alpha \) and \( \beta \), evaluations of multidimensional poverty are likely to vary according to the set of weights and final cut-off selected. As Duclos et al. (2006) note, approaches to multidimensional poverty measurement typically can be distinguished according to their respective emphasis on defining the poor as those deprived on one of many domains (the union approach) as opposed to those who define the poor as being deprived on all domains (the intersection approach). In turn, the AF class nests both these extremes as special cases – a pure intersection definition occurs when \( \kappa = 1 \); a pure union definition occurs when \( \kappa = \min(w_j) \).

The present aim is to use stochastic simulation methods to verify whether evaluations of changes in multi-dimensional deprivation over time are robust to alternative weights and cut-offs. I focus on the specific case of Mozambique, using data from three nationally-representative surveys (1996/97, 2002/03 and 2008/09); and where households are the microeconomic units of observation. Mozambique is not merely chosen for illustration – it has achieved one of the world’s most rapid and sustained rates of per capita economic growth since the end of conflict in 1992. However, recent consumption poverty estimates from the aforementioned surveys raise doubts whether aggregate growth has translated into broad-based welfare gains (DNEAP, 2010; Arndt et al., 2012b). Official estimates from the most recent rounds indicate that national headcount consumption poverty has stagnated at a little over 50% since 2002/03. Nonetheless, the same surveys indicate more consistent gains in non-consumption dimensions, including access to public goods such as education services. Consequently, it is appropriate to investigate what has happened to multidimensional poverty in Mozambique.

In order to apply the AF class of measures, it is necessary to select the deprivation dimensions to be included. This can be controversial; however, limitations such as the availability of consistent data (over time) as well as exclusion of highly correlated dimensions tends to limit feasible
Table 3: Dashboard of deprivation indicators from Mozambican household surveys

<table>
<thead>
<tr>
<th></th>
<th>1996/97</th>
<th>2002/03</th>
<th>2008/09</th>
<th>∆08–02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumption</td>
<td>68.3</td>
<td>54.2</td>
<td>54.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Non-food share</td>
<td>48.4</td>
<td>37.8</td>
<td>44.2</td>
<td>6.5</td>
</tr>
<tr>
<td>Shelter</td>
<td>74.2</td>
<td>72.0</td>
<td>66.8</td>
<td>-5.1</td>
</tr>
<tr>
<td>Potable water</td>
<td>91.1</td>
<td>85.3</td>
<td>80.4</td>
<td>-4.9</td>
</tr>
<tr>
<td>Economic assets</td>
<td>80.3</td>
<td>62.3</td>
<td>38.2</td>
<td>-24.0</td>
</tr>
<tr>
<td>Max. education</td>
<td>67.0</td>
<td>58.4</td>
<td>47.9</td>
<td>-10.5</td>
</tr>
<tr>
<td>Other assets</td>
<td>59.9</td>
<td>46.1</td>
<td>41.7</td>
<td>-4.4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>69.9</strong></td>
<td><strong>59.4</strong></td>
<td><strong>53.4</strong></td>
<td><strong>-6.0</strong></td>
</tr>
<tr>
<td>Observations</td>
<td>8,248</td>
<td>8,689</td>
<td>10,801</td>
<td>-</td>
</tr>
</tbody>
</table>

Source: own calculations.

Notes: each cell reports the share of households deprived on a given indicator based on deprivation thresholds discussed in the text; ‘Average’ is the column mean; the final column is the simple difference between the 2002/03 and 2008/09 values; ‘Observations’ is the sample size for each survey.

choices. The domains selected for the present exercise are summarised in Table 3, which is a dashboard of national deprivation headcounts for each domain in each survey round. The seven domains cover: consumption poverty (as per the official statistics, which yields a poverty line of under one US$ per person per day); the share of non-food items in household consumption (households are deprived if less than 30% of consumption goes to non-food items); access to shelter (a household is not deprived if they either have a toilet or their housing is made from durable material); access to clean water; ownership of economic assets (the household is not deprived if it has either a telephone, a car, a motorbike or a bicycle); ownership of other durable goods (the household is not deprived if has either a TV, a radio, an electric/gas cooker, or electric lights); and education (a household is deprived if no member has completed at least 5 years of schooling).

To run the stochastic simulation, I proceed in similar fashion to the previous example. Specifically: (i) I create a $7 \times 7$ matrix of equal weights ($\omega_j = 1/7$) and a corresponding vector of cut-off of values \{1/7, 2/7, ..., 1\}; (ii) using a shuffled Halton sequence, I draw a $9993 \times 7$ matrix of pseudo-random weights, each row of which is normalized to sum to one; (iii) for each row of the same matrix I draw $n$ elements, where $n$ is a random integer between 2 and 6, the sum of these elements corresponds to the chosen cut-off value; (iv) these two weight matrices are stacked and horizontally concatenated to the corresponding cut-off vectors yielding a $10000 \times 8$
Table 4: Estimates of sample coverage for AF headcount simulations

<table>
<thead>
<tr>
<th>Iteration</th>
<th>( U_{GS} )</th>
<th>( D_0 )</th>
<th>( U_{GS} )</th>
<th>( D_0 )</th>
<th>( U_{GS} )</th>
<th>( D_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.0</td>
<td>-</td>
<td>100.0</td>
<td>-</td>
<td>100.0</td>
<td>-</td>
</tr>
<tr>
<td>100</td>
<td>87.4</td>
<td>113.5</td>
<td>85.7</td>
<td>161.1</td>
<td>87.4</td>
<td>194.0</td>
</tr>
<tr>
<td>500</td>
<td>53.0</td>
<td>20.3</td>
<td>58.3</td>
<td>16.8</td>
<td>57.8</td>
<td>15.1</td>
</tr>
<tr>
<td>1000</td>
<td>35.8</td>
<td>7.5</td>
<td>38.9</td>
<td>4.7</td>
<td>37.9</td>
<td>6.0</td>
</tr>
<tr>
<td>2000</td>
<td>16.0</td>
<td>3.1</td>
<td>18.0</td>
<td>2.9</td>
<td>17.7</td>
<td>2.5</td>
</tr>
<tr>
<td>4000</td>
<td>8.7</td>
<td>2.1</td>
<td>8.0</td>
<td>2.7</td>
<td>7.1</td>
<td>2.8</td>
</tr>
<tr>
<td>6000</td>
<td>6.5</td>
<td>1.0</td>
<td>4.3</td>
<td>0.9</td>
<td>3.1</td>
<td>0.7</td>
</tr>
<tr>
<td>8000</td>
<td>4.4</td>
<td>1.5</td>
<td>3.4</td>
<td>1.2</td>
<td>2.2</td>
<td>1.2</td>
</tr>
<tr>
<td>10000</td>
<td>2.9</td>
<td>0.5</td>
<td>2.3</td>
<td>0.4</td>
<td>1.3</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Source: own calculations.

Notes: \( U_{GS} \) is the modified Good-Turing estimator due to Gandolfi and Sastri (2004), calculated after the number of iterations indicated in each row; \( D \) is a measure of distributional distance calculated according to a discrete version of equation (6), where the reference distribution is the outcome vector observed after the number of iterations indicated in the preceding row; \( D \) also uses a grid of 50 percentile points and sets \( \alpha = 0 \).

In presenting the results, it is sufficient for present purposes to concentrate on the (unadjusted) AF headcount measure. This follows Theorem 2(ii) in Alkire and Foster (2011), which shows that if welfare in group \( I \) first order dominates that of \( J \) for a given weight vector – i.e., \( I \)’s headcount is no higher than \( J \)’s over all cut-off points – then this directly implies dominance holds for the adjusted headcount. This result naturally extends to other weight vectors and, in the limit, the entire space of weights and cut-offs. Consequently, Table 4 reports coverage estimates for the vector of national AF headcounts, employing a rounding rule of three decimal places.

It is worth highlighting that, in principle, it is not necessary to apply a rounding rule when evaluating the robustness of the AF headcount measure. Equation (11) can be seen as a Boolean

\[ \text{matrix of inputs; (v) for each row of the input matrix I calculate the full range of AF measures for each round of the Mozambican surveys at the aggregate level and for three broad regions (North, Center, South); (vi) these simulated outcome vectors are then used as a basis for further analysis, including of simulation coverage.}^{11} \]

\[ \text{11For details of this specific algorithm, implemented in the R environment, are available on request from the author. Note that equal cut-offs are used initially because they provide broad coverage of the outcome space, including the supremum and infimum of the outcome measures.} \]

21
threshold function (see Crama and Hammer, 2011). Since the number of unique Boolean threshold functions is finite, it follows that the aggregate headcount also must take a finite number of values. In practice, however, it is not feasible to enumerate all possible Boolean threshold functions for a large number of dimensions. Thus a rounding rule remains useful; in this case it is again derived from equation (5), setting $\epsilon = .02$ and employing the lowest individual standard error estimated from the 10000 headcount iterations.

As before, the estimates of simulation coverage indicate relatively smooth asymptotic convergence toward zero as the number of iterations increases. A key difference, however, is that more than 6000 iterations are required for both measures to fall below 5% in all survey rounds. Even at 10000 iterations, the $U_{GS}$ measure suggests that the probability of seeing a ‘new’ headcount value on the next iteration remains greater than 1%. This reflects the greater complexity of the present problem, due to the larger number of input dimensions as well as reliance on microeconomic survey data. Despite this, it is evident that good coverage of the input space is highly feasible from a computational perspective. Also, the distance measures $D_0$ all indicate very small proportional changes to the quantile’s of the respective outcome distributions at the last block of iterations.

Using these results, pointwise comparisons of national headcounts reveal an unambiguous reduction in multi-dimensional poverty between 1996 and 2002, but more ambiguous performance between 2002 and 2008. This is illustrated in Figure 3, which plots empirical kernel density estimates of pointwise differences in AF headcount rates between adjacent surveys. It shows these differences are all greater than zero for the 1996-2002 comparison (with a minimum of 1.97 percentage points), implying that welfare was better in 2002 than in 1996 according to any choice of weights or cut-offs. Regarding the 2002-2008 comparison, the average of the vector of differences suggests a fall in multi-dimensional poverty of around 7.55 percentage points (from 57.16%) over the period. Nonetheless, for around 3% of iterations (points in the input space), the opposite conclusion is found. These increases in poverty are not negligible – the maximum is 6.07 percentage points, which is statistically highly significant. Thus, any proposition that multi-dimensional poverty has fallen between 2002 and 2008 is not strictly robust.

12 Counting the number of positive (monotone) Boolean functions of $n$ variables is known as Dedekind’s problem, for which no concise closed-form expression exists. A trivial upper bound is $2^{2^n}$, which is the number of all possible Boolean functions in $n$ variables. An active literature also seeks to enumerate all unique Boolean threshold functions, which are a sub-set of the monotone functions (see Korshunov and Shmulevich, 2002; de Keijzer et al., 2012; Kurz, 2012); however, no general solution has been found.

13 Using the approximation due to Kakwani (1993), the standard errors corresponding to the weights and cut-off that delivers this maximum are less than one quarter of a percentage point in both 2002 and 2008. In turn, this implies a t-statistic of around 17.73 on the difference in headcount estimates.
To get a more detailed sense of trends in welfare over time, Table 5 reports pointwise headcount comparisons for each pairing of location (region) and survey round. Similar to the presentation above, figures above the diagonal report the most favourable comparison between the row group (survey-region) and the column group, which is the supremum of the vector given by the column group’s headcount minus the row group’s headcount. Negative values indicate there is no sampled point in which the column group’s headcount is higher than the row’s. Values below the diagonal report the proportion of points over which the column group’s welfare is superior to the row’s (i.e., has lower headcount poverty).

Three points can be highlighted. First, households in the South of the country – which is economically dominated by the capital city – consistently have been the least deprived in a multi-dimensional sense on average. For example, households observed in the South in 1996/97 (‘S96’) welfare dominate households in the North and Center observed in 2002/03 on around
90% of all sampled input vectors. In fact, S96 welfare dominates that of households in the North and Center in 2008/09 on more than three quarters of all inputs. Second, the ambiguous picture of changes in multi-dimensional poverty reduction noted at the national level between 2002 and 2008 principally reflect developments in the Central region. Households in the Center observed in 2002 (‘C02’) welfare dominate households from the same region observed six years later on around 13% of input vectors. Third, whilst the Central region historically has shown higher multi-dimensional welfare than that of the North, this gap is closing. For instance, in 1996/97 Northern households (‘N96’) welfare-dominated Central households for less than 2% of input vectors. However, by 2008/09 the corresponding proportion had risen to 38% of sample points, suggesting that welfare comparisons between the North and Center are now much more sensitive to choices of weights and cut-offs.

### 6 Conclusion

This paper began by recognising the power of stochastic simulation methods to address analytically complex problems. One of these, the central focus of the paper, is the robustness of (high-dimensional non-linear) composite indicators to variations in constituent input parameters. A main aim was to address a weakness of simulation methods – namely, a lack of reliable ex post measures of the coverage of resulting simulated outcome distributions. This was considered
especially important in the case of robustness testing because, by definition, robustness refers to
the entire space of potential outcomes.

To address this gap, two measures of distributional coverage were proposed. The first was based
on the Good-Turing estimator developed during the Second World War. Refinements to this
estimator as well as its adaptation to contexts where outcomes are continuous, were discussed.
Second, an estimator of the proportional distance between two distributions (observed after
consecutive blocks of iterations) was suggested. As with the first estimator, in the limit this is
expected to converge to zero as the number of simulation iterations increases. Monitoring the
rate of convergence in empirical applications thus can be used to inform a simulation stopping
rule. Alternatively, estimates of ‘missing mass’ or ‘distributional distance’ after a fixed number
of iterations can be interpreted as indicating the degree of completeness (certainty) of the final
simulated distribution.

To provide guidance to practitioners, a generic algorithm combining these estimators with a
stochastic simulation was suggested. Illustrative empirical examples were then considered. The
first was a very simple macroeconomic application, based on the 2012 Human Development
Index (HDI). This showed that after 5000 iterations, the likelihood of observing a new index
value was less than 10% for all countries and less than 5% for the majority. In other words,
conclusions based on the actually-simulated outcome distributions could be asserted with at least
90% confidence. Analysis of these distributions revealed that country rankings are not uniformly
stable; rather, they depend on the degree of concordance across the three sub-indices aggregated
to give the final index. Additionally, a main advantage of employing stochastic methods to
simulate outcome distributions is that pointwise dominance comparisons can be made. These
pointed to nuanced insights regarding the robustness of HDI (rank) comparisons across selected
countries.

A second application was to the Alkire-Foster class of multi-dimensional deprivation measures.
Considered for the case of Mozambique, the convergence (completeness) estimates revealed
the simulated headcount distributions to be adequate after around 10000 iterations. Specifically,
taking a grid of fifty quantiles, the maximum proportional change of any one quantile occurring
between 8000 and 10000 iterations was less than 1.5%. The implication is that additional simula-
tions were unlikely to significantly add to our knowledge of the outcome distributions. Pointwise
comparisons of these headcount distributions indicated unambiguous welfare improvements
between 1996 and 2002; however, for around 3% of iterations welfare was deemed to have fallen
between 2002 and 2008. Further analysis revealed this was driven by dynamics in the Central
region.

In sum, when combined with estimates of simulation coverage, stochastic methods provide a simple and ‘almost certain’ means to undertake robustness tests of complex empirical questions. A core advantage, relative to other approaches, is that such simulations directly give an estimate of the empirical distribution of outcomes. With these distributions in hand, a wide range of additional analysis, including pointwise comparisons can be made. Analytical extensions, not considered here, might include quantifying the impact of marginal changes in input parameters on chosen outcomes; other options include using more sophisticated inferential methods to verify dominance conditions.

References


