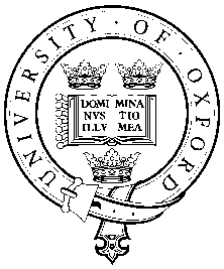
Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford




**OPHI WORKING PAPER NO. 115**

# The Measurement Properties of Multidimensional Poverty Indices for Children: Lessons and Ways Forward

Martin C. Evans* and Akmal Abdurazakov**

March 2018

**Abstract**
This paper considers the measurement properties of indices used to measure multidimensional child poverty in the developing world. Two indices are considered in detail: the Alkire Foster method (Alkire & Foster 2010) and the 'categorical counting' method as exemplified by UNICEF poverty indices based on methodologies by Gordon et al. (2003) and De Neubourg et al. (2013). This analysis examines the underlying differences between the two methodologies in two stages. First, using hypothetical data we consider the differences in measurement properties that arise from the axiomatic construction of indices using a laboratory approach. Second, we use harmonized Demographic and Health Surveys data from three countries to examine how the properties found in the laboratory data lead to actual differences in the measurement of the prevalence of multidimensional poverty within and across countries, and the ability of indices to monitor changes in the prevalence of multidimensional poverty. The paper concludes by considering the findings from the analysis and how they could be taken forward in future measurements of poverty prevalence and reduction in Sustainable Development Goals targets and indicators.

**Keywords:** Child Poverty, Multidimensional Poverty, Poverty Measurement

**JEL classification:** I32, J13

* Senior Research Fellow, Overseas Development Institute, and Research Associate, Oxford Poverty and Human Development Initiative (OPHI), University of Oxford. m.evans@odi.org.uk.

** Independent Consultant.

This study has been prepared within the OPHI theme on multidimensional measurement.

ISSN 2040-8188 ISBN 978-19-1229-103-8

**Acknowledgements**
The authors are indebted to a large set of colleagues for discussions with and comments on this and earlier versions of the paper. In particular, our thanks go to Dave Gordon and colleagues for allowing access to harmonized DHS datasets at University of Bristol; and to Dominic Richardson, Conchita D'Ambrosio, Brian Nolan, Stephen Jenkins, Sabina Alkire, Adriana Coconi, Ana Vaz, Elina Scheja, Pablo Suarez Becerra, Chris Oldiges, Paula Ballon and Maria Ana Lugo; members of the Multidimensional Poverty Peer Network; attendees at the ISI conference in Marrakech in July 2017, especially discussants, Milorad Kovacevic and José Cuesta; and HDRO and UNDP staff at the New York presentation of the paper in November 2017. Our special thanks go to Attila Hancioglu who helped in the earlier conference draft with text and comments on MICS surveys. Of course, any errors of calculation or interpretation remain our responsibility.

**Citation:** Evans, M.C. and Abdurazakov, A. (2018). 'The measurement properties of multidimensional poverty indices for children: lessons and ways forward'. OPHI Working Paper 115, University of Oxford.

*The Oxford Poverty and Human Development Initiative (OPHI) is a research centre within the Oxford Department of International Development, Queen Elizabeth House, at the University of Oxford. Led by Sabina Alkire, OPHI aspires to build and advance a more systematic methodological and economic framework for reducing multidimensional poverty, grounded in people's experiences and values.*

This publication is copyright, however it may be reproduced without fee for teaching or non-profit purposes, but not for resale. Formal permission is required for all such uses, and will normally be granted immediately. For copying in any other circumstances, or for re-use in other publications, or for translation or adaptation, prior written permission must be obtained from OPHI and may be subject to a fee.

Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK
Tel. +44 (0)1865 271915 Fax +44 (0)1865 281801
ophi@qeh.ox.ac.uk http://www.ophi.org.uk

The views expressed in this publication are those of the author(s). Publication does not imply endorsement by OPHI or the University of Oxford, nor by the sponsors, of any of the views expressed.

## 1. Introduction

The inclusion of multidimensional poverty in the 2015 Sustainable Development Goals (SDGs) is a long-awaited recognition of its importance and relevance, but it has raised new requirements for measurement. Additionally, the poverty goals and targets in the SDGs also prioritize children, making child multi-dimensional poverty a key area for measurement of SDG progress. Prior to the SDGs, children's multidimensional poverty indices were constructed and used for advocacy and other purposes, but less importance was placed on their properties as poverty measurement tools. To establish robust baselines and capture changes in poverty over time, clear cardinal and scalar properties will be optimal to set and monitor SDG targets. This paper considers the underlying methodologies of approaches that are in place to meet this challenge.

Measurement of multidimensional poverty has grown rapidly since UNICEF's ground-breaking 2003 report on global multidimensional poverty for children (Gordon et al. 2003). Other methods appeared, most notably the Multidimensional Poverty Index (MPI), which is based on the Alkire Foster (AF) methodology (Alkire and Foster 2011) and was adopted by the United Nations Development Programme (UNDP) for their Human Development Reports beginning in 2010 (UNDP 2010). The AF methodology (but not MPI) has now also been adopted by the World Bank in implementing an index that fulfills a recommendation of the 'Atkinson' report (Commission on Global Poverty 2017). In Europe, Eurostat has developed multiple material deprivation measures that also consider non-monetary measures of material poverty using a multidimensional approach (Eurostat 2015).

The literature on multidimensional indices and measurement has also expanded exponentially in the past 10–15 years across a wide set of disciplinary and technical approaches: from indices developed in theoretical terms using mathematical and econometric specifications in economics journals to descriptive and normative studies using qualitative and quantitative data in policy and children's journals. However, one characteristic of the literature is that multidimensional child poverty has been dominated by the latter approaches and thus has mostly avoided the technical scrutiny of econometricians in the former.

Our paper proceeds as follows. The remainder of this introductory section first outlines the household survey data used for the indices and then the multidimensional index methodologies that we compare. The analytical part of the paper follows in two sections: Section 2 compares 'laboratory' tests of the main indices to a simple benchmark sum-count index; Section 3 considers the indices as implemented in actual household survey data in three countries in order to assess how far the laboratory findings are present in real data. The paper then reviews its findings and offers conclusions regarding the implementation of SDG poverty measurement.

### 1.1 Household Survey Data

We concentrate on the main sources of survey data that are currently feeding into large scale multidimensional child poverty measurement: USAID-supported Demographic and Health Surveys (DHS) and the UNICEF-supported Multiple Indicator Cluster Surveys (MICS).

Two key aspects of MICS and DHS data influence the performance of individual-level multidimensional poverty indices for children: the levels at which data are collected and age-specific differences in data.

First, surveys collect data at two levels, individual and household, that are used together in indices. .Indicators at the household level – such as water, sanitation, dwelling construction, occupation, and assets – cover all children in a household. If these indicators are positive, all children who live in those households are deprived and data used to create the index is clustered at the household level. Additionally, and by contrast, data at the individual child level – on their health, nutrition, education, and other areas of child and maternal wellbeing is collected at the individual level. These indicators will vary at the individual level and identify children who are deprived and not deprived in any household. Indices thus report a 'count' of deprivations across both levels. However, index scores will be potentially dominated by household-level clustered indicators and any analysis of differences at the individual level will be biased by the balance of household-level to individual-level indicators within the index. This leads to a limited ability to capture individual-level differences in the composite index.

Second, there are age-specific differences in observable indicators at the individual level. Data on children is collected specifically for certain age-related risk groups. For instance, detailed anthropometric data is only collected for those aged less than 60 months while school enrolment or attendance is only observed in school-aged children. This means that indicators for nutrition, health, education, and other crucial areas of child poverty and wellbeing are not available for all ages of children. This creates 'censored' data at the individual level and further limits the assessment of individual-level differences across all ages of children. When such censored data is joined with household-level data in indices, it reinforces measurement problems and further complicates the identification of individual-level differences, which may already be obscured by clustered data in households. The issue of age- or population-specific data and its effect on multidimensional indices was discussed at length by Dotter and Klasen (2015) and led to revisions in the UNDP's MPI (Kovasevic and Calderon 2016).

It is worth remembering that neither DHS nor MICS were set up with the goal of creating a balanced multi-indicator index. And, they certainly were not designed *ex-ante* to capture multidimensional poverty from an individual perspective – the method often used for multidimensional child poverty profiles in developing countries.

### 1.2 Multidimensional Poverty Indices

We limit our analysis and review to counting indices drawn from DHS and MICS survey data. These indices arithmetically sum indicators that have been reduced to binary form (deprived or not deprived). A counting index in its simplest form can be the sum of each indicator expressed in binary form. Thus, the sum count index for each child has values from zero to ten. We use such a simple counterfactual sum count as a baseline comparison index in our analysis.

The two indices in our analysis differ in the way in which they assign indicators to dimensions and in the allocation of weights to indicators and/or dimensions.

*Alkire Foster*. Using the Alkire Foster (AF) methodology results in an index calculated by the sum of indicators that has no axiomatic assumptions on weighting of either indicators or dimensions. However, the most well-known version of the approach, the global MPI, is often wrongly assumed to be the axiomatic expression of the methodology. In practice, indicator-level weights have been determined according to the dimensions that contain the indicators. Using such an approach, the sum of dimensions and indicators will be one, but each indicator can be weighted independently or per share allotted to the dimension of which it is a part. In the global MPI, three dimensions were set to reflect the three dimensions of UNDP's Human Development Index (health, education, and living standards), with each having an equal weight of 1/3. The education and health dimensions have two indicators each with resulting weights of 1/6; the living standards dimension has six indicators, resulting in indicators weights of 1/18. But, while the MPI often dominates the discussion, it should be considered an early and seminal variant of the AF methodology, not its essential representation. For instance, Vietnam's multidimensional poverty measure (MOLISA 2016) has five dimensions and ten indicators; thus, each indicator has a weight of 1/10 and is an exact replication of the simple sum-count index discussed earlier. Indeed, national MPIs adopted by governments across the developing world differ from the global version in many ways. MPIs have, to date, often been specified at the household level. Recent individual-level AF indices include Klasen and Lahoti (2016) specifying an individual-level version of the global MPI. Children's multidimensional poverty can be captured by disaggregating household-level MPI by age – as recently done at the global level for the first time by Alkire et al. (2017). But specific child-level AF measures, while established early in the literature (Roche 2013), have been much later arriving in practice in national poverty profiles. Bhutan was the first country to officially adopt an individual-level 'child MPI' (Alkire et al. 2016) while more are currently underway in Panama, Vietnam, Maldives, Afghanistan, and other countries.

*Categorical Counting (CC).* This term is ours and refers to a number of indices that use a normative rights-based approach.[1] The crucial arithmetic differences from both the AF and sum-count approaches are fourfold. First, the dimensions are counted in order to produce the index score. Second, the aggregation of indictors into dimensions uses the 'Boolean' logic of the 'union approach', meaning that the dimensional binary score is one if *any* of the indicators in that dimension is positive, or is zero if not one of them is positive. Third, there is no necessity for a consistent number of indicators per dimension. Some dimensions contain a single indicator (often 'sanitation' and 'water' dimensions in practice), while others can contain two or more indicators. And, finally, it is axiomatic that each dimension has equal weight due to a normative rights-based assertion used in the approach.

These indices more longstanding – starting with Gordon et al.'s 2003 global child poverty profile and followed by UNICEF's global poverty and disparities profiles in around 50 countries, as well as a specific version of that index for Latin America (CEPAL/UNICEF 2010, 2012). More recently, Multiple Overlapping Deprivation Analysis (MODA) was developed using a similar approach (De Neuburgh et al. 2013).

The prevalence of multidimensional poverty using these indices depends in part on two assumptions regarding the level of indicator cutoffs for deriving the binary deprivation indicators and the composite score that results from the summing of those indicators. Versions of the indices discussed have been set to both more extreme indicator cutoffs and to differing thresholds (poverty lines) of the index scores.

The seminal innovation arising from the AF methodology was to replicate the set of core of Foster-Greer-Thorbecke (FGT) decomposable poverty metrics (Foster, Greer, and Thorbecke 1984) for poverty headcount, poverty gaps, and poverty intensity. This enabled the AF methodology to meet most of the axiomatic requirements of poverty measurement established in the poverty literature (Alkire et al. 2015).

### 1.3 The Analysis

There are a few studies that directly compare these indices in practice. For instance, there are comparisons of individual-level MODA indices with MPI household-level indices (Calderon & Evans 2015, Hjelm et al. 2016). These studies find differences in the level and composition of poverty, but such differences are difficult to interpret when it is not clear if such differences are due to the underlying methodology, the construction of indices and indicators in survey data, the use of different indicators, differences in the construction of similar indicators, or from underlying data cleaning work such as the trimming of data for outliers, etc. The global MPI's early years were characterized by criticism of multidimensional assumptions and measurement outcomes from poverty economists established in the monetary approach (Ravallion

1 Another term could be 'dimensional counting', but as AF decomposition often produces results by dimension, we consider our term less ambiguous.

2011 and others). Over time, the body of analytical literature about the AF approach has grown much larger and now includes comparisons to statistical and econometric measurement practices of poverty. For instance, tests of robustness and sensitivity for MPI have been undertaken (Alkire 2014, for instance) and alternative theoretical measurement approaches compared (Rippin 2010 and others). Technical evaluation of the CC approach has been minimal by comparison.

Our primary research questions go back to the applied question of poverty measurement for the SDGs: How do the AF and CC approaches perform in terms of three clear questions? 1. How do they differ in cardinal and scalar properties? 2. How do they set robust baselines? 3. How do they assess changes to poverty prevalence or intensity over time to meet SDG targets?

Our approach is to consider the axiomatic measurement properties of the indices in the first instance and then to assess how such identified properties affect performance in actual household survey data. We break from the typical economic literature on measurement properties by not basing our analysis on algebraic proofs. The reasons for this are several. First, we want to reach a wider audience that is concerned with the practical applications of these indices instead of just the readership of highly technical econometric, mathematical, and statistical journals. Second, algebra can sometimes be a 'black box' that hides uncertainty and assumptions. For instance, when considering counting indices, the Greek symbol sigma $\Sigma$ is a crucial algebraic element that may represent a cumulative sum of different and non-consistent components; a consistent Greek symbol may obscure differences between the ordinal, categorical, and cardinal nature of the components to be summed. Third, our worked examples from laboratory data benefit from iteration and include the outcomes of a trial and error process in some instances. Indeed, one important lesson from our work is that demonstrating ex-ante theoretical measurement problems in the lab can identify unforeseen measurement properties – such as the potential of indices to saturate.

We end our analysis by moving from laboratory data to implement indices using real survey data and thus move from discussions of pure measurement theory into applied practice. We test real micro-survey data from three countries to see if the findings from laboratory tests are validated.

Our motivation is to go beyond the simple descriptive comparisons of indices already in place and concentrate on the underlying methodology and its applied outcomes for data and poverty profiles. But the potential lure of specific 'brands' of indices (MODA, MPI) is strong, and thus we avoid brand comparisons of named indices but instead concentrate on the underlying measurement approaches and their assumptions. Our comparisons and review are of AF and CC methodologies and their generalizable properties. We recognize the investments that have gone into named indices, but our analysis is of the AF and CC approaches and not solely of the indices that spring from them: MODA, MPI, CEPAL/UNICEF, etc.

## 2. Tests Using Laboratory Data

We construct laboratory data for 10,000 hypothetical observations with ten non-specified binary indicators for each observation. In our first set of tests we randomly (coin-flip) allocate a score of one or zero to each indicator. This means that in our first set of estimates all indicators are independent of each other and there is, by definition, no correlation – an assumption that we change later when we reflect further on our sensitivity and monotonicity analysis. For the random data, we make comparisons between indices from 100 Monte Carlo trials using coin-flip random allocation to provide robust estimates at the 99.9% level. This provides us with a baseline distribution and result for a mean score of 0.5 across the ten indicators against which to compare index performance.

We have three test indices.

1. AF index: Our test index uses the same weights as those in the global MPI – reflecting both the most well-known version of their methodology and the individual-level MPI index demonstrated by Klasen and Lahoti (2016). Four of the ten indicators have weights of 1/6 and six indicators have weights of 1/18.
2. CC index: Our test index uses weights and approaches from the applied literature (Gordon et al. 2003 and De Neuburg et al. 2013). There are five dimensions. Three dimensions are populated by two indicators in a union approach; one dimension has three indicators in a union approach; and one dimension has just one indicator.
3. Sum-count index: This is used for comparison and baselining purposes. There are ten indicators, each with an equal weight of 0.1.

### 2.1 Tests for Index Scales and Baselines

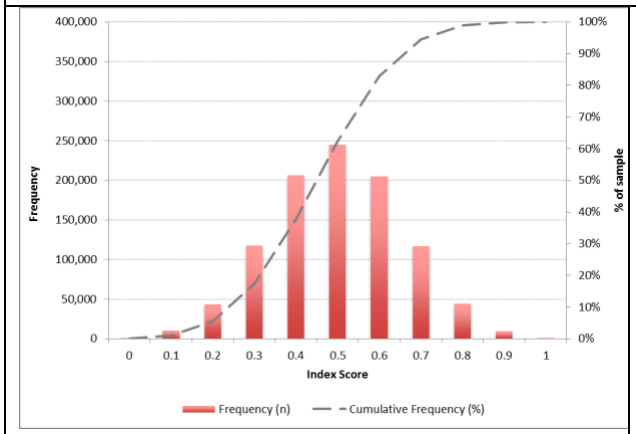
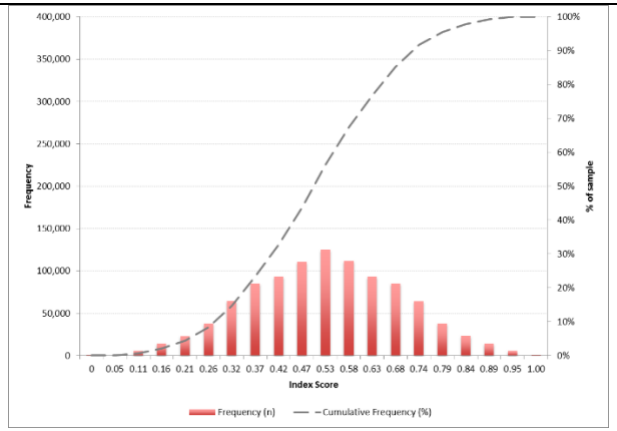
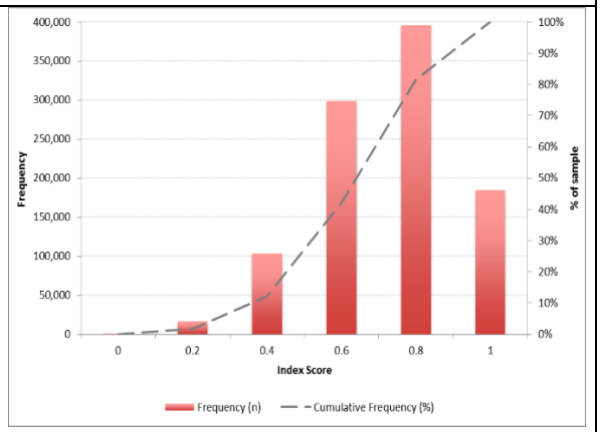
Figure 1 shows the results. Figure 1a shows the sum-count benchmark index, which, by definition and construction, produces a mean and median score of 0.5. It has a normal distribution that has 11 scalar increments of 0.1 between zero and one, with a standard deviation of 0.16. These are bench-test reference results to compare against the other indices.

Figure 1b shows the AF index also results in a bell curve distribution (but Shapiro-Wilk tests do *not confirm* it is a normal distribution) with 18 equal increments of 0.566 (the value of the smallest weight) between zero and one. Having index weights smaller than 1/10 results in a more granular distribution compared to the sum-count benchmark but with a consistent mean and median at 0.5 and a standard deviation of 0.18.

Figure 1c shows the results for the CC index, which stand in stark contrast to both the benchmark and AF specifications for the same data and underlying distribution of deprivation scores. The first thing to

**Figure 1: Baseline Results**

Monte Carlo Simulations (100 trials)

| **a) Sum Count – Benchmark** | **b) Alkire Foster** | **c) Categorical Counting** |
|---|---|---|
|  |  |  |



**Properties of Index Scales**

Values and Increments between 0 and 1 (indices all normalized to 1)

| 10 increments – each 0.1 | | 18 increments – each 0.057 | | 5 increments – each 0.2 | |
|---|---|---|---|---|---|
| **Summary Statistical Properties** | | | | | |
| *Mean* | 0.50 | *Mean* | 0.50 | *Mean* | 0.73 |
| *Median* | 0.50 | *Median* | 0.50 | *Median* | 0.80 |
| *Std Deviation* | 0.16 | *Std Deviation* | 0.18 | *Std Deviation* | 0.19 |
| *Skewness* | 0.00 | *Skewness* | 0.00 | *Skewness* | -0.41 |
| *Kurtosis* | 2.80 | *Kurtosis* | 2.62 | *Kurtosis* | 2.81 |

note is the far less granular distribution, as counting dimensions rather than summing indicator scores results in only five increments of 0.2 between zero and one on the scale. But most noticeably, the distribution for the CC approach is greatly skewed and has higher index scores overall with a mean of 0.73 and a median of 0.8. It is worth repeating that these results come from the same underlying set of observations and prevalence of indicators. Simply said, the CC specification exaggerates the prevalence of multidimensional poverty (which can be read as a headcount at any threshold score from 0.2 to 1) compared to the other indices. This finding confirms the discussion and findings of Chakravarty and D'Ambrosio (2006), Rippin (2010), and others who found that a union approach results in an exaggeration of poverty. This is our first finding from the laboratory work and is important for our second question: How do the indices set robust baselines? The inherent characteristic of exaggeration for CC versus the other indices is a property that must be explicitly addressed when assessing such baselines in practice.

Of course, we are mindful that particular specifications of AF or CC indices may give different results. However, our detailed laboratory work suggests that different iterations (weights or assignment of deprivations to dimensions) *do not alter the fundamental findings of difference* or the finding that the CC methodology exaggerates poverty at all thresholds compared to the AF and sum-count specifications in any form. Confirmatory results can be obtained from the authors.

### 2.2 Tests for Sensitivity and Monotonicity

Our findings so far on the shape and properties of the different distributions formed by the indices also raise points that are relevant to our third question for poverty measurement: How do indices indicate if poverty is changing over time to meet SDG targets?. This is a key question for both accurately identifying differences between subgroups and for tracking changes over time.

Figure 2 shows the results from Monte Carlo trials of repeated incremental changes of plus and minus 10 percentage points for a randomly selected indicator across all three indices. There are two main findings of interest: the level of change, which will be affected by the weight of the indicator that is changed, and the consistency of change for positive and negative values, or symmetry. Figure 2 shows that the sum-count index, with each indicator having a weight of 0.1, has a symmetrical profile of change from 0.45, where indicator prevalence is zero, to 0.55, where indicator prevalence is 100%, from a starting point of 0.5. The AF index has a similar symmetrical profile but produces larger changes in index scores for the same incremental change in indicator prevalence:.0.42 overall, if prevalence is reduced to zero, and 0.58 overall, if prevalence is increased to 100%, from the same starting point of 0.5. On the other hand, the CC index changes asymmetrically from its much higher mean point of 0.72. Decreasing prevalence in one indicator reduces the score by 0.04 points to 0.68, but increasing indicator prevalence to 100% raises the

score by 0.06 points, and the difference between these points is statistically significant at 99% using t-tests. This suggests that the CC index is asymmetric.

**Figure 2: Changes in Index Scores from Incremental Change to Any Indicator**

100 Monte Carlo Trials

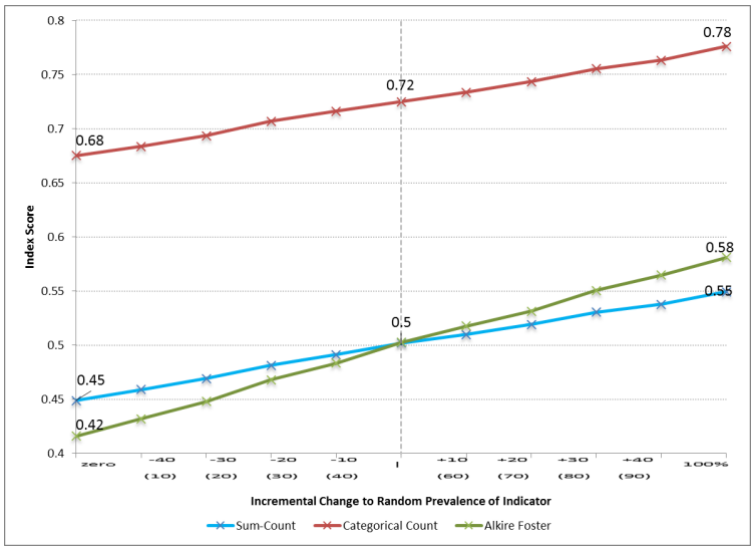
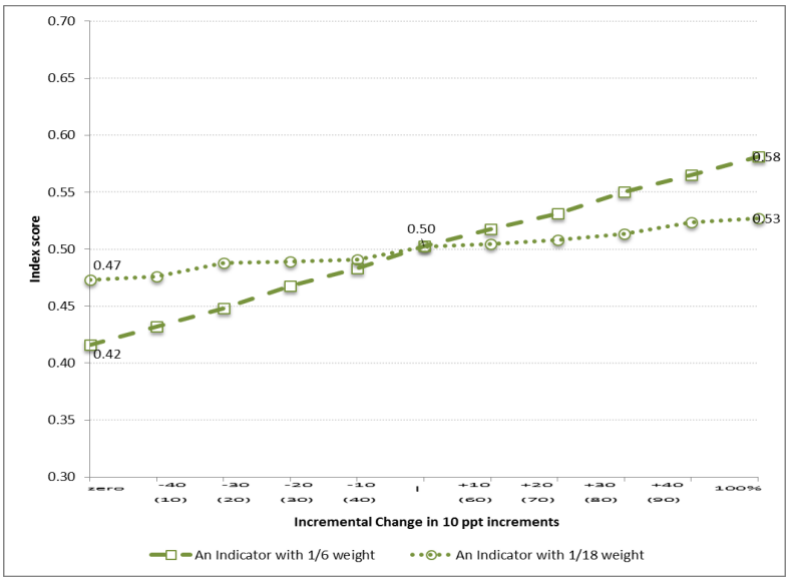



Figure 3 shows the results of incremental change to indicators that have differential weights in the AF index. The level of incremental change differs: large change is the result of a higher indicator weight but changes are similarly symmetric across incremental change to indicators of differential weight.

**Figure 3: AF: Incremental Change in Indicator by Assigned Weight**

100 Monte Carlo trials




When we attempt to replicate Figure 3 for the CC index, we are hindered by the design of our laboratory data used to this point and the specific characteristics of the index. Our first set of laboratory data set all indicator prevalence to 0.5, with random, independent (non-correlated), indicators. Our lab tests using a consistent 0.5 prevalence were inconsistent and the reasons for this only became apparent once we

established by trial and error that the CC index suffered from saturation effects, which prevent the index from *consistently* increasing or decreasing overall scores due to underlying increased or decreased prevalence.

**Figure 4: CC: Incremental Change in Indicator by Union Assumption with 20%, 50% and 80% Random Prevalence**

100 Monte Carlo Trials

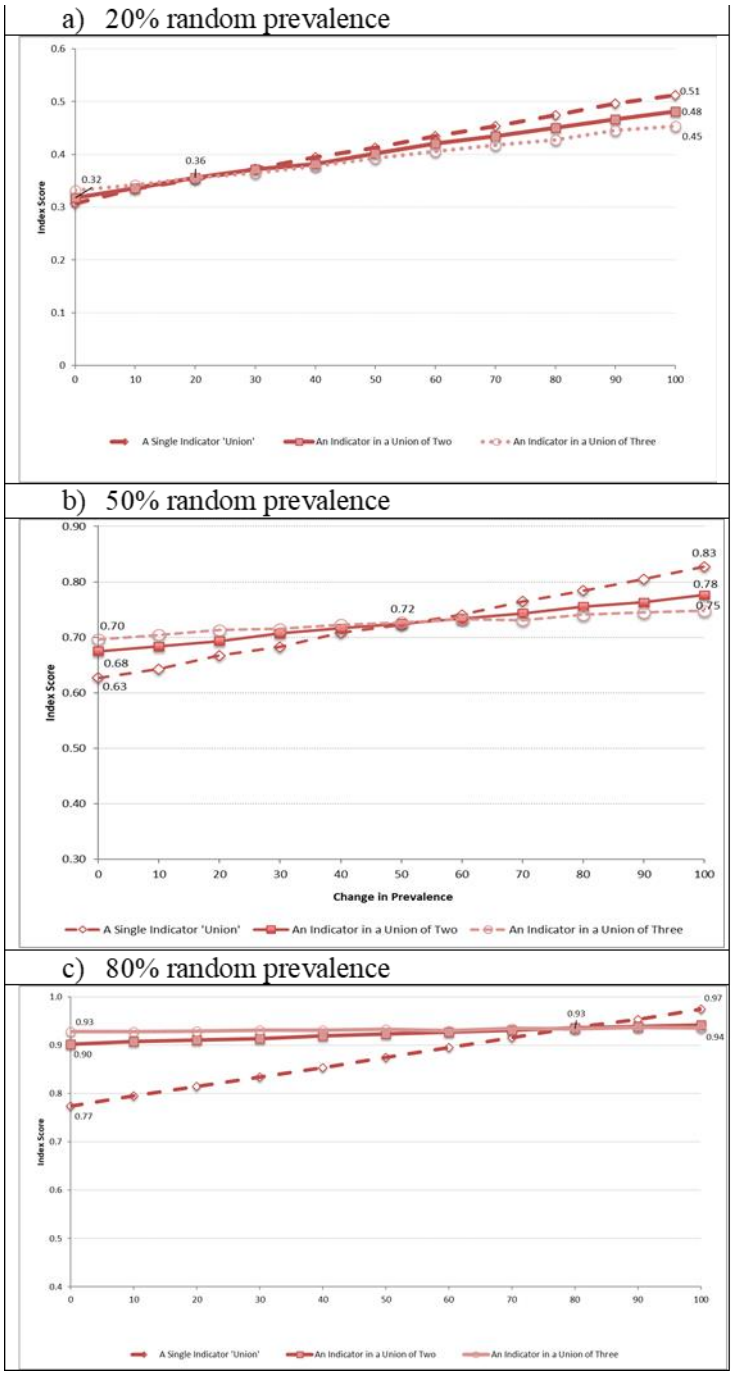



Figure 4 shows the effects of saturation as well as the underlying patterns of asymmetry that are seen in the CC index. We show the incremental effects on the index of changing indicators according to their aggregation assumptions (whether they are in a single, double, or triple union). We use the same random

assignment and non-correlated approach but change the lab data to produce 20% and 80% average score outcomes, alongside our original 50% prevalence simulation.

The results show clear asymmetry but with asymmetric attributes that are not consistent. The results differ both in levels of saturation and by the 'union assumption' for the indicator used to simulate incremental change. At low levels of saturation (20% random prevalence) the index converges towards the 0%, while at high levels of saturation (80% random prevalence) the index converges towards 100% prevalence. This overall asymmetry is the result of differing asymmetry for the indicators according to their union with other indicators. As expected, and by design, the indicator that has no other indicators in union ('single union') has the clearest monotonic characteristics across all levels of saturation. However, the indicators in union with one or two other indicators clearly show differential slopes and much smaller levels of incremental change over all the ranges of incremental change of indicator. They also become 'flat' at high levels of concentration – where there is little if any change to the overall score from incremental changes in indicator prevalence. These characteristics of the CC index are thus axiomatic by design and show that inconsistent asymmetry occurs due to both differences in the union assumption of the indicator and in the level of saturation. The conclusion for monotonicity is thus pretty clear.

### 2.3 Testing the Effects of Correlation

So far, we have relied on independent randomly assigned prevalence across our ten hypothetical indicators, but, if we change our assumption of independence we can assess if correlation between indicators affects the comparison of indices. The CC index will inherently rely, in part, on linked probabilities for indicators that are assessed in union. Such linked probabilities will not affect indicators that are not aggregated in union – both for the sum-count and AF indices and for single indicator categorical dimensions in CC indices.

Figure 5 illustrates the effects of negative and positive correlation using the union approach employed by CC indices. It shows that positive and negative correlations produce inconsistent arithmetic sums when the union aggregation approach is employed. The observed prevalence of indicator A is 50% and B is 10% – with a negative correlation resulting from B containing observations that are not contained in A. This leads to a 60% combined union count. In the alternative case, C has 20% prevalence and is positively correlated as both observations are common between C and A. The result of the union aggregation of A and C is 50%. This suggests a fundamental measurement problem for monotonic index performance as a change in indicator prevalence will not produce increases in the overall index score due to positive and negative correlation differences.

**Figure 5: Differential Outcomes of Negative and Positive Correlation in Union Aggregation**

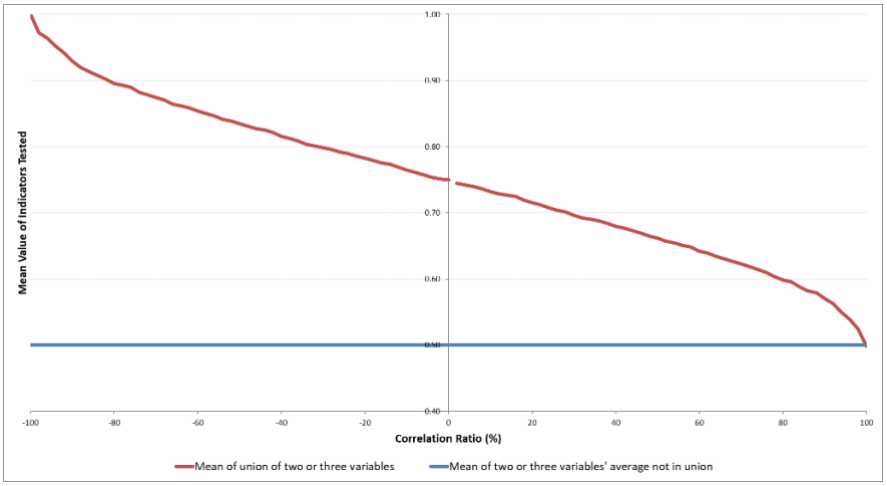
| obs | A | B | | Union A & B | C | | Union A&C |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | x | | | x | x | | x |
| 3 | x | | | x | x | | x |
| 4 | | | | | | | |
| 5 | x | | | x | | | x |
| 6 | - | x | | x | | | |
| 7 | x | | | x | | | x |
| 8 | x | | | x | | | x |
| 9 | | | | | | | |
| 10 | | | | | | | |
| | 50% | 10% | negative correlation | 60% | 20% | positive correlation | 50% |

To demonstrate the effect of correlation on index monotonicity, we return to using laboratory data, and to do so we reformulate our source data of 10,000 observations to drop the assumption of independent randomly assigned indicators and use correlation ratios between pairs or triads of indicators. To have high levels of statistical certainty for our estimates we use 100 fixed intervals of the correlation ratio, representing 2 percentage point increments between negative 100 and positive 100% correlation. Using these data we test the relationship for linearity between indicators using regression and are able to report results that have 0.01 p values.

Figure 6 shows two tests that assess how the level of correlation affects indices. Figure 6a shows results for pairs and triads of indicators, and Figure 6b shows the results of correlations of pairs on the overall index scores. We set the underlying average indicator prevalence for the data to 50% prevalence for both of these tests.

**Figure 6: The Effect of Correlation on Multidimensional Counting Indices**

**6a. On Mean Value of Pairs/Triads of Indicators**



**6b. On Index Scores from Correlation of Indicator Pairs**

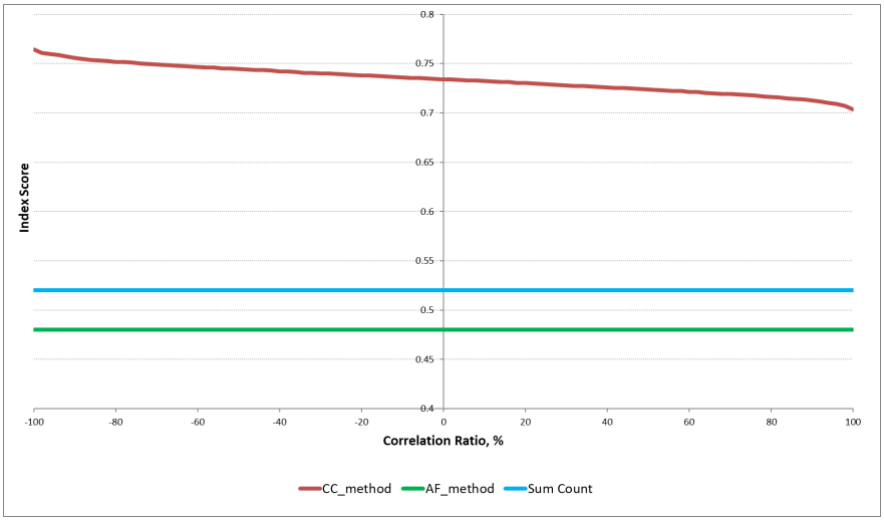
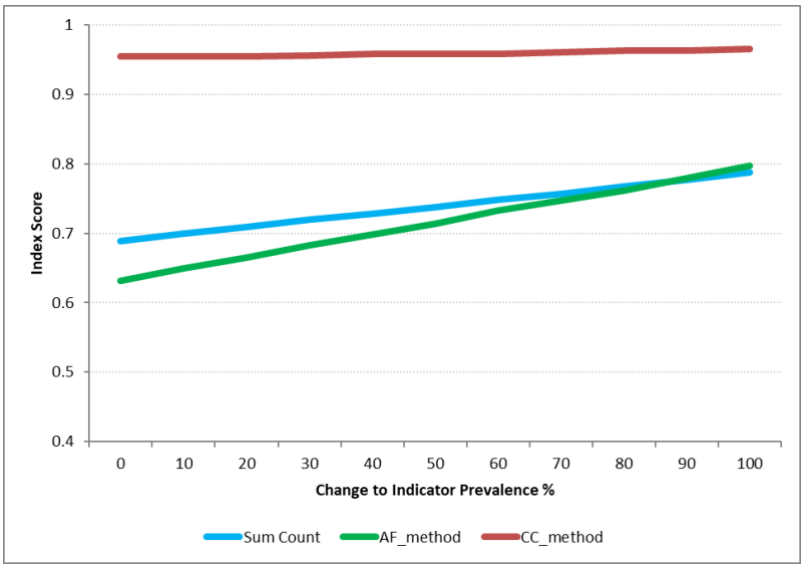



Figure 6a shows the effect of union aggregation of the mean value of pairs or triads of indicators. The red line shows the changing mean of indicators that are aggregated using the union approach – either as pairs (twofold union) or triads (threefold union). The blue line shows the mean for two or three indicators not aggregated in union. We see that only indicators in union are affected by correlation, and that differences in correlation are non-linear in their effect on the mean values of those indicators. Figure 6b follows up on this finding to show the results of different levels of correlation of pairs of indicators on overall index score across the three indices, using the same pair of indicators used in Figure 5a at (as before) 50% prevalence. We thus test to see how the indices react to positive correlation between two indicators out of all ten indicators. Figure 6b demonstrates how the CC index score is influenced non-linearly by the level of correlation between those two indicators. Neither the AF nor sum-count indices are influenced by differing correlation levels between that identical pair of indicators.

Our final test then takes these results and applies them to the three indices to assess the effect of high correlation between the pair of indicators on differences in indicator prevalence – the test for monotonicity that we used in the previous section. Figure 7 shows the results. Correlation is seen to make the CC method non-monotonic as there is no change in index score from underlying changes to indicator prevalence when indicators are subject to strong leading indicator correlation. This is not seen in the sum-count or AF index at the same level or correlation for the same changes in prevalence for such correlated indicators.

**Figure 7: Changing Index Outcomes from High Correlation of Indicators**




## 2.4 Household Clustering, Age-Specific Censoring

We return to our original laboratory dataset of 10,000 observations and ten randomly assigned independent deprivations to consider the differences that occur as a result of the two measurement problems we identified earlier when discussing survey data and individual-level child poverty indices: (1) indicators can be clustered at the household level while others are at the individual level and (2) individual-level indicators are age specific and any observation (individual) not in that age range is missing and thus censored for that indicator. We address these two issues in turn.

How do the original distributions described in Figure 1 change if some of the indicators are at the household level and thus have the same value for all members of that household? We reconfigure the laboratory dataset to assign individual observations to 'pseudo-households' randomly. We then regenerate the dataset with the following approach:

1. We set the average number of observations per pseudo-household as three;

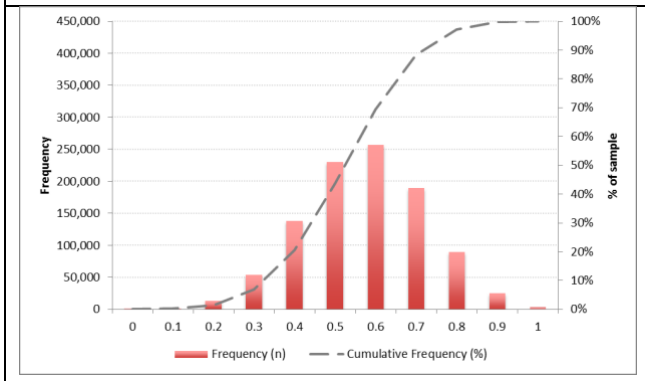
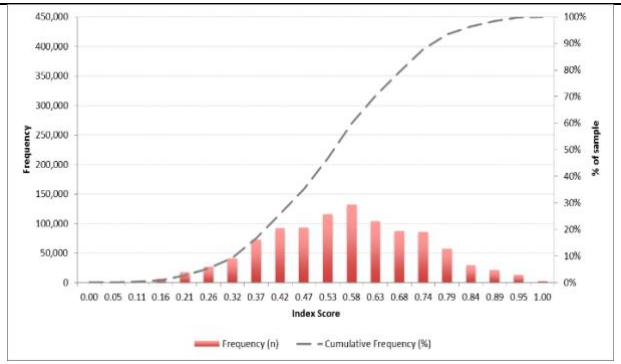
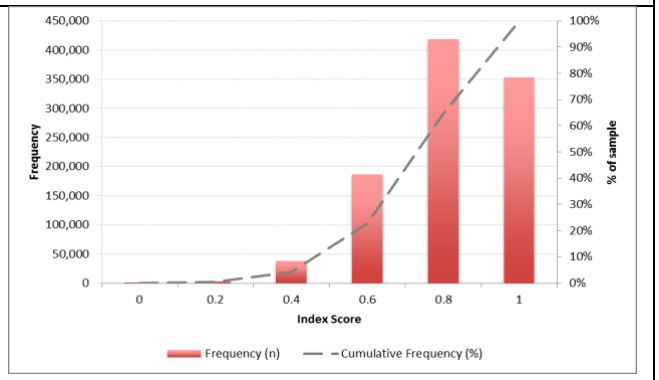
2. We allow up to a maximum of seven observations to be allocated to a random household;

3. To this new distribution, two indicators were assigned to be household-level indicators; the remaining eight indicators remain at the individual level and are thus not clustered.

These assumptions are designed to illustrate the potential effects of household clustering compared to a random allocation and are not intended to be representative of actual household formation and size. We tested other specifications and found that the index reaction to clustering was unchanged in nature.

**Figure 8 Revised Results Allowing for Household Clustering (2/10 indicators at household level)**

Monte Carlo Simulations (100 trials)

| a) Sum-Count Benchmark | b) Alkire Foster[2] | c) Categorical Counting[3] |
|---|---|---|
|  |  |  |

**Summary Statistical Properties and Differences from Baseline**

| Revised Statistic | Difference from Baseline | Revised Statistic | Difference from Baseline | Revised Statistic | Difference from Baseline |
|---|---|---|---|---|---|
| *Mean*. . . . . 0.57 | +0.07 ** | *Mean*. . . . 0.54 | +0.04 ** | *Mean*. . . . . 0.82 | +0.09 ** |
| *Median*. . . . 0.60 | +0.10 | *Median*. . . 0.56 | +0.06 | Median. . . . . 0.8 | 0 |
| *Std Deviation*. . 0.15 | -0.01. n.s. | *Std Deviation*. 0.18 | -0.002 n.s. | *Std Deviation*. . .0.170 | -0.022.n.s. |
| | | | | | |
| *Skewness*.-0.06 | | *Skewness* -0.01 | | *Skewness* -0.68 | |
| *Kurtosis*. 2.83 | | *Kurtosis* 2.61 | | *Kurtosis* 3.03 | |

2 Dimension 1(I;I), Dimension 2(I;I) and Dimension 3(I;I;I; I;HH;HH) (HH= household-level indicator)

3 Dimension 1(I;I) Dimension 2(I;I), Dimension 3(I;I), Dimension 4(I; I;HH) and Dimension 5(HH)

The results are shown in Figure 8. Household clustering increases the skewness of all three distributions compared to their baselines in Figure 1, but the degree and effect of skew varies by index. The CC index increases skew most and increases its mean score from 0.73 to 0.82. Household clustering thus seems to increase the potential of the CC index to saturate as discussed earlier. On the other hand, the sum-count and AF indices are leaner functions, and the inclusion of household-level observations produces an upward shift in the poverty line as individually differing indicators are replaced with repeated household-level versions. The mean score rises from 0.50 to 0.57 for the benchmark, but from 0.50 to 0.54 for AF, a reflection of the fact that, on average, lower weighted indicators were affected. This raises the possibility that AF- type differential weights could be used to address household clustering effects on individual-level indices. We return to this point later in our discussion.

But our test on the impact of household clustering on two out of ten indicators is not indicative of what happens in practice with child poverty indices, where we see that household- level indicators are a much higher proportion of all indicators – often six and sometimes eight out of ten indicators are at the household level (Gordon et al. 2003, de Neuburgh et al. 2013).

Table 1 gives an indication of the additional differences that would result from higher levels of household clustering – with six out of ten indicators at the household level. The CC index's saturation is now clear, with a mean of 0.93 and median of 1. These results suggest a clear caveat for using indices of this type in contexts with high levels of deprivation and a high proportion of indicators at the household level. On the other hand, the AF index under these same assumptions maintains its lower means and medians and skewness compared to the sum-count benchmark, another indication that differential indicator weighting for household-level variables is worth considering in applied index work.

**Table 1: Household Clustering at Higher Margin (6/10 indicators household level)**

| | **Sum-count** | **AF**[4] | **CC** [5] |
|---|---|---|---|
| Mean | 0.69 | 0.66 | 0.93 |
| Median | 0.7 | 0.67 | 1 |
| Standard deviation | 0.14 | 0.16 | 0.13 |
| Skewness | -0.52 | -0.34 | -1.82 |
| Kurtosis | 3.45 | 3.19 | 6.41 |

4 Dimension1(I;HH), Dimension 2(I;HH) and Dimension 3(I;I;HH; HH;HH;HH).

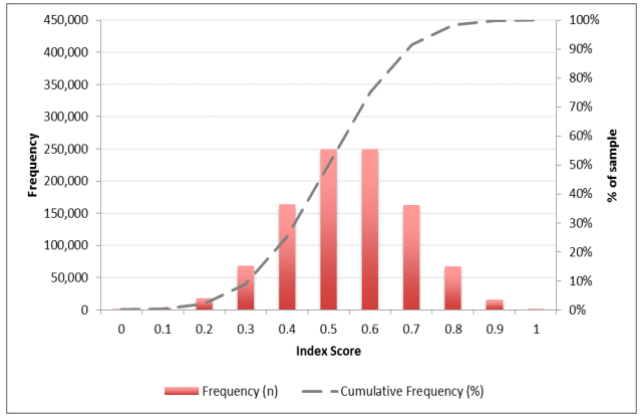
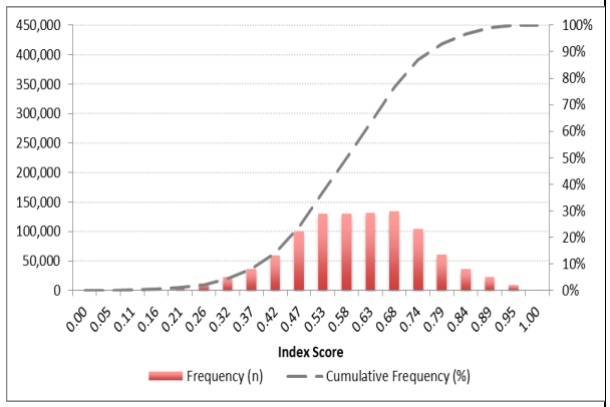
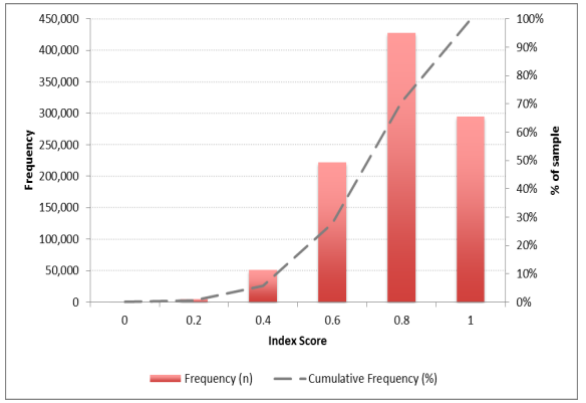
5 Dimension 1(I;HH) Dimension 2(I;HH), Dimension 3(I;HH), Dimension 4(I; HH;HH) and Dimension 5(HH).

We now turn to look at age-specific censoring of individual-level indicators and its potential impact on results from our first set of randomly assigned laboratory data. By definition, only individual-level variables can be age specific. If individual-level indicators are censored, the ratio of observed individual-level variance falls relative to common individual indicators and to the household-level repeated values in any index. As a result, the potential for the dominance from clustered household-level variables rises for any index. To capture this effect, we further transform the household clustered version of our laboratory dataset that was seen in Figure 7. This allows us to consider the cumulative and marginal changes from censored data alongside clustered data. We censor prevalence for a *single indicator* from 10,000 to 7,100 to illustrate the underlying population size of the zero- to four-year-old population in the three countries (UNDESA 2015) that we consider later in section 3. To do this we replace all positive one values with zero for these 2,900 cases. While in reality, these values would be 'missing', we chose to replace with zeros to avoid the need for to adjust to underlying population differences. However, in the applied use of indices, population reweighting would require more consideration and we discuss that issue below.

Figure 9 gives the revised results for this transformation and its effect on the indices. We show the net change in summary statistics compared to those seen from household clustering alone in Figure 8. The results show that histograms for the benchmark and AF-MPI have flattened curves around mean values. For the sum-count benchmark this leads to a small but statistically significant (at 1%) reduction to the mean compared to the Figure 7 results. For AF there is a small statistically significant increase (at 1%) in the mean. Both these specifications show reduced standard deviations that result from fewer positive values for indicators. The CC specification shows a small significant (at 1%) decrease in mean score and an increase in the standard deviation. However, underlying skewness measures have risen to the highest level across all three simulations in Figures 1, 8, and 9. These results show the CC index reacts differently to changes in individual-level prevalence compared to the sum-count and AF indices and this will further contribute to problems of monotonicity.

**Figure 9: Results Allowing for Household Clustering and Age-Specific Incidence of Deprivation**

Monte Carlo Simulations (100 trials)

| **Sum Count – Benchmark** | **Alkire Foster** | **Categorical Counting** |
|---|---|---|
|  |  |  |



**Summary Statistical Properties and Differences from Iteration Shown in Figure 8**

| Revised Statistic | Difference from Figure 8 | Revised Statistic | Difference from Figure 8 | Revised Statistic | Difference from Figure 8 |
|---|---|---|---|---|---|
| *Mean*. . . . 0.55 | -0.025 ** | *Mean*. . . . 0.58 | 0.039 ** | *Mean*. . . . 0.79 | -0.025 ** |
| *Median*. . . 0.52 | -0.082 | *Median*. . . 0.58 | 0.0272 | *Median*. . . 0.8 | 0 |
| *Std Deviation*. 0.15 | -0.002.n.s. | *Std Deviation*. 0.16 | -0.023 n.s. | *Std Deviation*. 0.18 | 0.005.n.s. |
| | | | | | |
| Skewness -0.04 | | -0.17 | | -0.56 | |
| Kurtosis 2.84 | | 2.97 | | 2.90 | |

Notes: ** significant at 1% using two tailed t-test

The results from the alternative allocation of household-level indicators to dimensions are available from the authors but do not alter interpretation of results from this example.

We now turn to a final consideration of the potential effects of clustering and censoring by looking at the potential effects on correlation. For individual-child-level indices, the issue of correlation is obviously partly determined by what is common to all children in a household and what is particular to children of a certain age. How does clustering and censoring affect the assessment of indicator correlation and how can our indices minimize the effects of correlation bias?

**Figure 10: Correlation Tests and Clustered and Censored Indicators**

**Data from 5 Households**

| hhd | person | *a* | *b* | *c* | *d* | *e* | *f* | *g* | *h* | *i* | *j* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 4 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 4 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 5 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 5 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |

**Effects on Correlation Ratios**

- Changes sign when censored cells change from zero to missing
- Changes sign when household variables reweighted
- Changes sign when both adjustments made

**i) Baseline: No adjustment**

| | a | b | c | d | e | f | g | h | i | j |
|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 1 | | | | | | | | | |
| *b* | 0.151 | 1 | | | | | | | | |
| *c* | 1 | 0.168 | 1 | | | | | | | |
| *d* | 0.647 | 0.270 | 0.649 | 1 | | | | | | |
| *e* | -0.034 | -0.352 | -0.171 | -1 | 1 | | | | | |
| *f* | -0.537 | -0.157 | -0.651 | 0.011 | -0.006 | 1 | | | | |
| *g* | 0.091 | 0.687 | 0.100 | -0.541 | -0.408 | -0.057 | 1 | | | |
| *h* | 0.826 | -0.157 | 0.893 | 0.200 | -0.408 | -0.057 | 0.619 | 1 | | |
| *i* | -0.582 | 0.628 | -0.695 | 0.130 | -1 | 0.551 | 0.692 | -0.139 | 1 | |
| *j* | 0.346 | -0.195 | 0.213 | -0.577 | 0.043 | 0.196 | 0.785 | 0.785 | 0.139 | 1 |

**ii) Adjustment for censoring (0 values changed to missing in blue cells above)**

| | *a* | *b* | *c* | *d* | *e* | *f* | *g* | *h* | *i* | *j* |
|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 1 | | | | | | | | | |
| *b* | -0.210 | 1 | | | | | | | | |
| *c* | 1 | -0.327 | 1 | | | | | | | |
| *d* | 1 | 0.293 | 0.777 | 1 | | | | | | |
| *e* | -1 | -1 | -1 | -1 | 1 | | | | | |
| *f* | -0.210 | 0.293 | -0.327 | 0.293 | -1 | 1 | | | | |
| *g* | -0.284 | 0.424 | -0.401 | -0.746 | -1 | 0.424 | 1 | | | |
| *h* | 1 | -0.746 | 1 | 0.182 | -1 | 0.424 | 0.307 | 1 | | |
| *i* | -0.641 | 1 | -0.754 | -0.135 | -1 | 1 | 1 | -0.320 | 1 | |
| *j* | -1 | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 |

**iii) Adjustment for clustering (1/n) for household-clustered variables in green cells above)**

| | *a* | *b* | *c* | *d* | *e* | *f* | *g* | *h* | *i* | *j* |
|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 1 | | | | | | | | | |
| *b* | -1 | 1 | | | | | | | | |
| *c* | 1 | -1 | 1 | | | | | | | |
| *d* | 1 | -1 | 1 | 1 | | | | | | |
| *e* | -1 | 1 | -1 | -1 | 1 | | | | | |
| *f* | -1 | -1 | -1 | 0 | 0 | 1 | | | | |
| *g* | -0.544 | -1 | -0.544 | -0.601 | 0.601 | 0.121 | 1 | | | |
| *h* | 0.336 | -1 | 0.336 | 0.293 | -0.293 | -0.091 | 0.619 | 1 | | |
| *i* | -0.475 | -1 | -0.475 | 0.037 | -0.037 | 0.765 | 0.692 | -0.139 | 1 | |
| *j* | -0.253 | 1 | -0.253 | -0.616 | 0.616 | -0.346 | 0.785 | 0.785 | 0.139 | 1 |

**iv) Both Adjustments: ii) and iii) compared to i)**

| | *a* | *b* | *c* | *d* | *e* | *f* | *g* | *h* | *i* | *j* |
|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 1 | | | | | | | | | |
| *b* | -1 | 1 | | | | | | | | |
| *c* | 1 | -1 | 1 | | | | | | | |
| *d* | 1 | -1 | 1 | 1 | | | | | | |
| *e* | -1 | 1 | -1 | -1 | 1 | | | | | |
| *f* | -1 | -1 | -1 | 0.144 | -0.144 | 1 | | | | |
| *g* | -1 | -1 | -1 | -0.814 | 0.814 | 1 | 1 | | | |
| *h* | -0.052 | -1 | -0.052 | 0.307 | -0.31 | 0.503 | 0.307 | 1 | | |
| *i* | -1 | -1 | -1 | -0.320 | 0.320 | 0.773 | 1 | -0.320 | 1 | |
| *j* | -1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 | 1 |

In Figure 10 we visually demonstrate how an assessment of correlation is affected by clustering and censoring. We show hypothetical data (not randomly assigned) from the first five households of a hypothetical dataset with ten indicators. The results reflect tetrachoric correlation tests in accordance with the binary ordinal data in these indices. The first correlation matrix shows the results for the original data as a baseline. The second matrix shows the comparison with the first matrix and thus the effect on the correlation ratios of censoring those data shown in blue in the original matrix, representing values originally designated as 'zeros' but changed to missing values. The effect of this adjustment is to change the sign of correlation ratios in the nine instances shaded blue. The third matrix shows the results of reweighting the values of clustered household-level indicators using a simple per-capita approach (for a household with four observations we reweight the indicator value from 1 to 0.25 (¼), and so on). Again, this adjustment

changes the sign of the correlation ratio in 13 cases, shaded green. The fourth and final matrix uses both adjustments, and the resulting changes in sign for the correlations are confirmed in 15 cells, which are highlighted orange. This test is designed to visually show the potential effects of censoring and clustering on correlation ratios and is illustrative only; the small samples in our demonstration mean that we put no weight at all on the changes in value of correlation ratios or on the robustness of changes in sign.[6]

What effects will these changes in correlation resulting from controlling for clustering and censoring have on the indices? The changes in sign that come from censoring and clustering adjustments mean that CC indices will be at most risk for resulting bias, as we have seen how different signs of correlation between indicators have the ability to alter monotonicity the most in that index. But CC indices, due to their axiomatic normative adoption of rights-based equal weights, do not allow changing index weighting assumptions to empirically reflect the effects of clustering or censoring. The implementation of these indices is not consistent, but the more recent MODA approach (de Neuburgh et al. 2013) has produced different indices for age-specific subgroups of children, unlike earlier CC approaches (Gordon et al. 2003, UNICEF/CEPAL 2009). On the other hand, AF approaches are more adaptable to empirically derived indicator weights that can adjust for over- and under-representation of indicators. But the use of frequency weights to adjust for censoring needs to be better addressed across all types of indicators, as suggested in the review of the global MPI (Klasen and Dotter 2014, Kovacevic and Calderon 2015).

## 3. Indices Using Household Survey Data

In this section, we test some of the key findings from the laboratory work using real survey data. At this point it is crucial to restate that this work will not replicate the actual indices that are in place. We are testing the underlying methodologies, not comparing the 'branded indices' – such as MPI and MODA – of those methodologies.

We use harmonized survey data prepared by Professor D. Gordon and his team at the University of Bristol using DHS and MICS surveys. This ensures that consistent approaches to indicator specification and data cleaning have been used across all the different national datasets. Our data comes from three 2010 DHS surveys for Colombia, Bangladesh, and Tanzania. To avoid some of the problems of age-specific censoring and in the interest of space and concision, we limit our indices to the population aged less than five years old. We take ten indicators and construct three indices from those indicators: the sum-count benchmark

6 Using a larger laboratory dataset or survey data of significant sample size showed similar changes to the size and sign of correlation ratios, results are available from the authors.

index, the AF index with three dimensions (2, 3, and 5 indicators per dimension), and the CC index with five dimensions (2, 3,1,1,3 indicators per dimension).

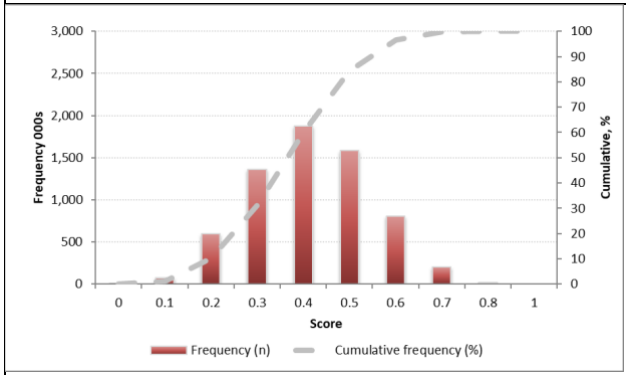
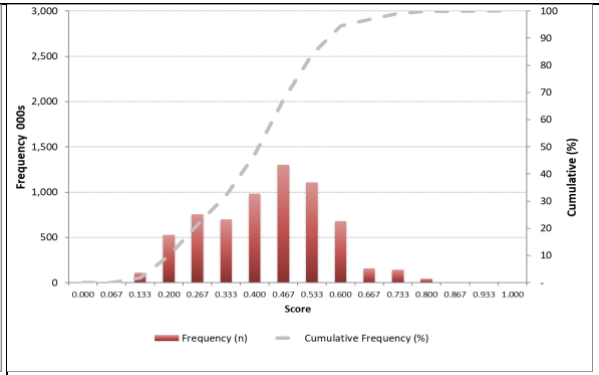
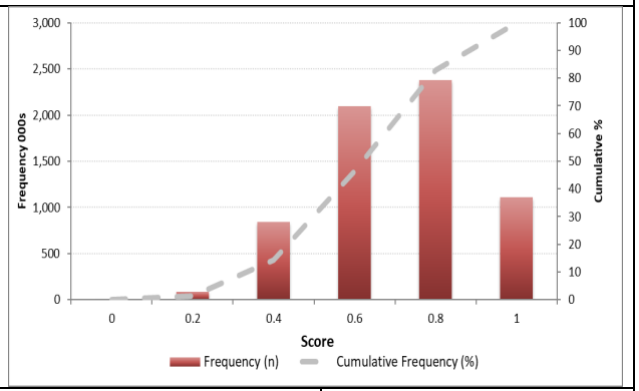
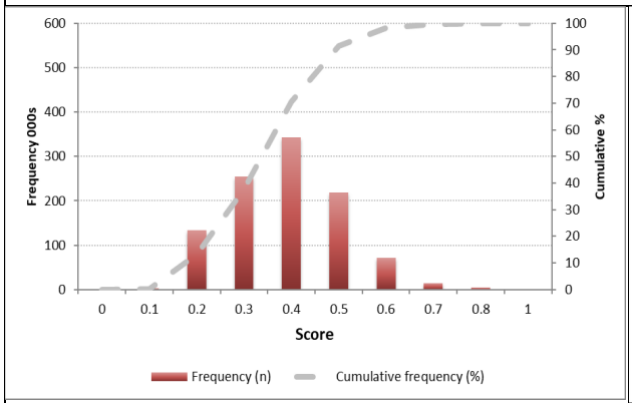
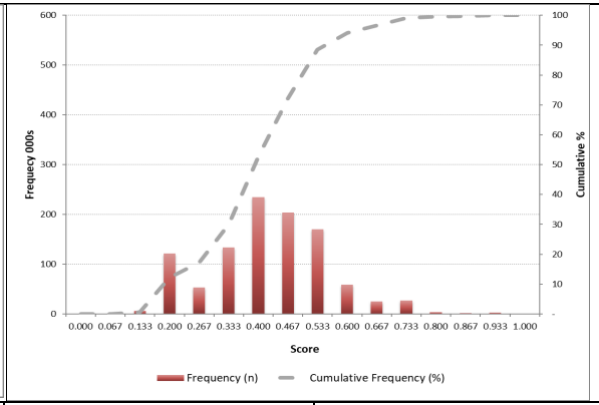
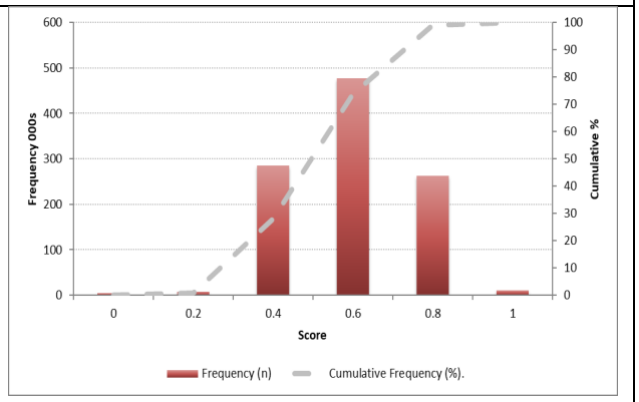
**Table 2: Indicators, Dimensions, and Weights for Test Indices**

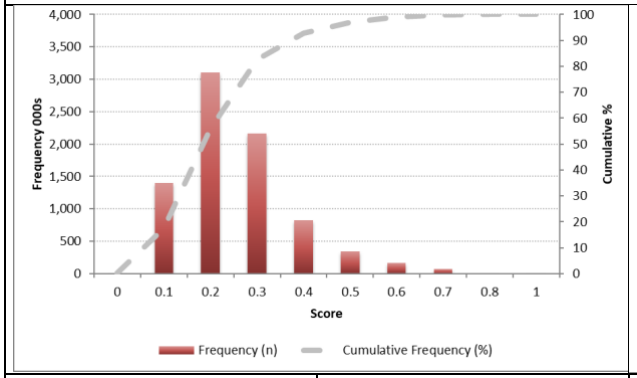
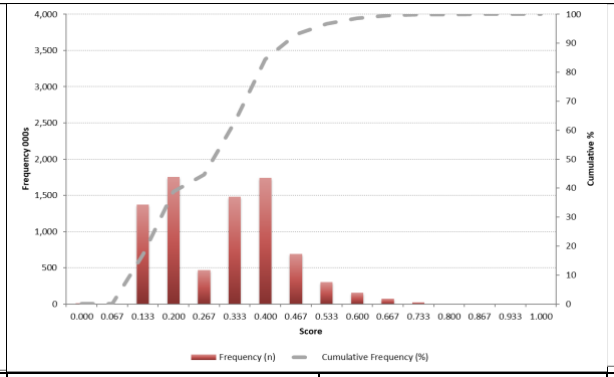
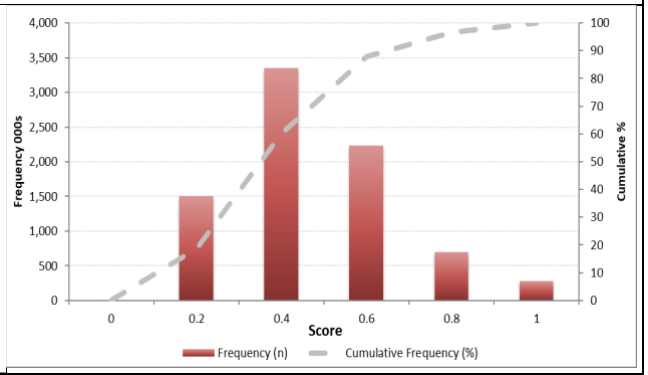
| **Categorical Counting: Composition of Dimensions** | | | | | |
|---|---|---|---|---|---|
| **Nutrition** | Infant feeding | Wasting | | | |
| **Health** | DPT all | Unskilled birth attendant | Child mortality | | |
| **Water** | Drinking water | | | | |
| **Sanitation** | Toilet type | | | | |
| **Living Standards** | Overcrowding | Wealth, low quintile | Info devices | | |
| | | | | | |
| **AF method: Composition of Dimensions** | | | | | |
| **Nutrition** | Infant feeding | Wasting | | | |
| **Health** | DPT all | Unskilled birth attendant | Child mortality | | |
| **Living Standards** | Drinking water | Toilet type | Overcrowding | Wealth, low quintile | Info devices |

Our indices are not designed to be relevant or inherently robust or meaningful in and of themselves because our motivation is not to design and test an optimal index. We do not test our set of indicators for their suitability, reliability, validity, or underlying robustness in their performance for any overall index specification because we are not interested in how these indices accurately assess multidimensional poverty for this test, only in their comparative performance according to underlying measurement properties. Our choice of indicators is also dictated by our desire not to replicate an actual index already in place. We have chosen some indicators that are used in multidimensional indexes and others, such as 'lowest quintile of wealth index' that are not, and, perhaps, never should be.

Figure 11 shows the key results for each of the three indices in each of the three countries, and we limit the reporting of results to simple comparisons of means and maxima in order to establish whether the lab results we saw earlier that indicate an exaggeration of poverty continue to be seen between CC and the other indices using actual survey data. The sum-count and AF indices all provide similar mean headcounts and similar maxima; using these combinations of deprivations, the maximum score is 0.8 across all countries for the sum-count index and 0.87 for the AF index. However, the CC index gives consistently higher mean scores compared to the other indices – around 50% higher. Additionally, the maximum score for a CC index is always 1.00, representing the outcome from counting dimensions, or categories of

**Figure 11: Headline Results for Indices in Three Counties: Distributions, Means, and Maximum Values**

| a) Sum Count | | b) Alkire Foster. | | c) Categorical Counting | |
|---|---|---|---|---|---|
| **TANZANIA** | | | | | |
|  | |  | |  | |
| *Mean* | 0.42 | *Mean* | 0.40 | *Mean* | 0.72 |
| *Maximum* | 0.8.0 | *Maximum* | 0.87 | *Median* | 1.00 |
| **BANGLADESH** | | | | | |
|  | |  | |  | |
| *Mean* | 0.40 | *Mean* | 0.40 | *Mean* | 0.61 |
| *Maximum* | 0.80 | *Maximum* | 0.87 | *Maximum* | 1.00 |

| **COLOMBIA** | | | | | |
|---|---|---|---|---|---|
|  | |  | |  | |
| *Mean* | 0.23 | *Mean* | 0.25 | *Mean* | 0.44 |
| *Maximum* | 0.80 | *Maximum* | 0.87 | *Maximum* | 1.0 |

deprivation, rather than sums of the underlying deprivation indicators. These findings support the earlier laboratory work on both cardinal and scalar properties and confirm the findings of exaggeration for CC indices versus AF and sum count-indices.

How do the indices perform when considering monotonicity? We show tests for two indicators: water and the presence of a skilled birth attendant. We have chosen these as they reflect the marginal cases and are illustrative of the underlying measurement properties we examined in the laboratory data.

Water is a household-level variable that has a great implicit weight in a CC index because it is a single-variable dimension, but has a lower implicit weight in the other indices.

The presence of an unskilled birth attendant is an individual-level variable that has a low implicit weight in a CC index because it is in a union of three indicators in a single dimension, whereas in the other indices it is measured using its indicator prevalence with slightly differing weights.

**Figure 12: Sensitivity Tests for Water and Presence of Skilled Birth Attendant Indicators**

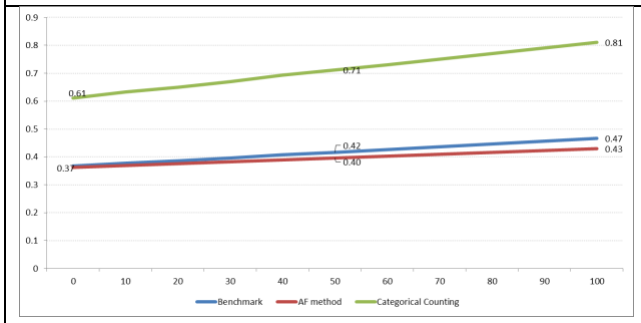
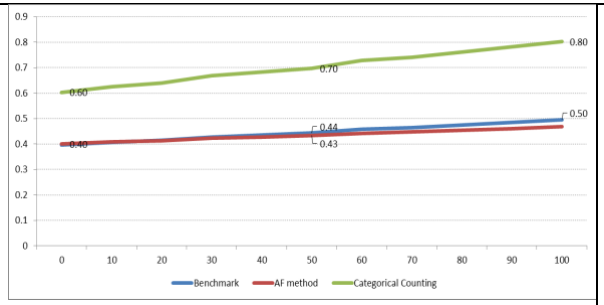
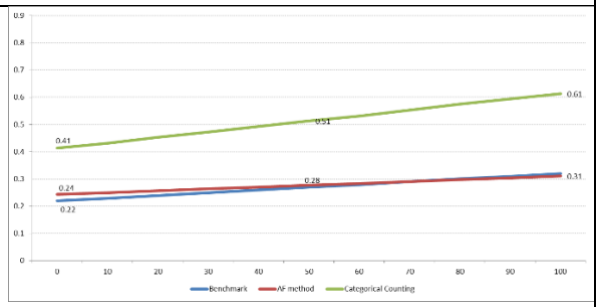
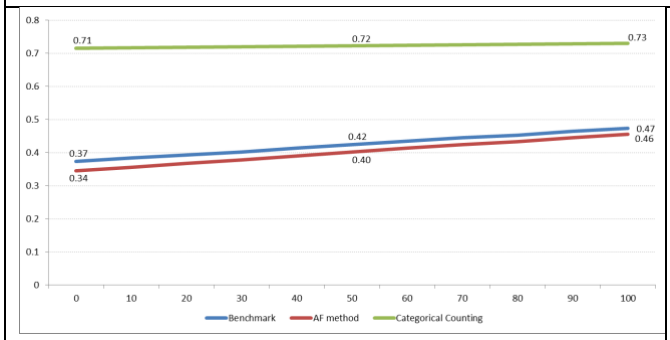
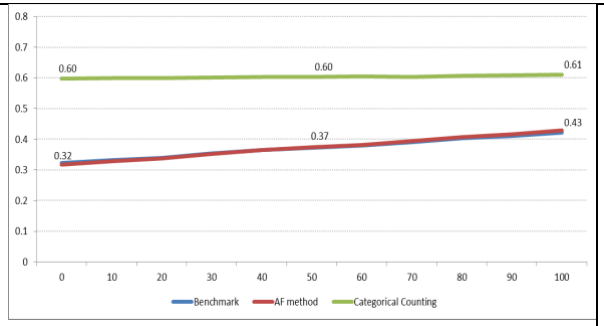
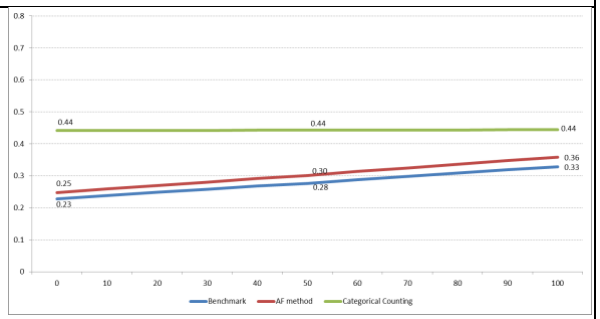
| TANZANIA | BANGLADESH | COLOMBIA |
|---|---|---|
| Water | | |
|  |  |  |
| Presence of Skilled Attendant at Birth | | |
|  |  |  |

Figure 12 shows common results across indices and across countries in the changes to both indicators from incremental changes in prevalence. Incremental changes to the prevalence of the water indicator have a big impact on the CC index across all countries, as represented by its status in a single-variable dimension. The change in overall slope from zero to 100% prevalence is much steeper when compared to the sum-count and AF indices and the absolute changes in index values are far higher overall. On the other hand, changes to the prevalence of the skilled birth attendant has little, if any, discernible effect on the CC index – as shown by the flat line in Figure 12; however, for the other indices, changed prevalence is clearly associated with increased or decreased incremental index scores. These results confirm what we saw in the monotonicity tests for the laboratory data but are more clearly interpretable for applied poverty measurement and for planning poverty reduction. Any investment by governments in the provision of additional services could result in inconsistent and arbitrary changes in poverty: giving rise to either large or no change in prevalence using the CC method.

**Table 3: Tanzania: Regional Ranking for Multidimensional Poverty Index Scores**
**Multidimensionally Poor Children Under 5**

| Regions | Sum Count Score | Rankings | | |
|---|---|---|---|---|
| | | Sum Count | Alkire Foster | Categorical Count |
| Tabora | 0.486 | 1 | 1 | 1 |
| Rukwa | 0.467 | 2 | 3 | 9 |
| Shinyanga | 0.465 | 3 | 2 | 2 |
| Mara | 0.464 | 4 | 4 | 3 |
| Tanga | 0.455 | 5 | 6 | 4 |
| Mwanza | 0.447 | 6 | 8 | 5 |
| Singida | 0.440 | 7 | 9 | 6 |
| Kigoma | 0.437 | 8 | 7 | 11 |
| Manyara | 0.436 | 9 | 11 | 12 |
| Dodoma | 0.435 | 10 | 10 | 10 |
| Lindi | 0.430 | 11 | 12 | 8 |
| Pemba North | 0.430 | 12 | 5 | 15 |
| Mbeya | 0.421 | 13 | 14 | 13 |
| Kagera | 0.420 | 14 | 13 | 7 |
| Pemba South | 0.393 | 15 | 15 | 18 |
| Arusha | 0.390 | 16 | 17 | 21 |
| Pwani | 0.386 | 17 | 20 | 14 |
| Mtwara | 0.384 | 18 | 19 | 16 |
| Morogoro | 0.379 | 19 | 18 | 19 |
| Zanzibar North | 0.373 | 20 | 16 | 24 |
| Iringa | 0.359 | 21 | 22 | 17 |
| Zanzibar South | 0.353 | 22 | 21 | 25 |
| Ruvuma | 0.347 | 23 | 23 | 20 |
| Kilimanjaro | 0.337 | 24 | 26 | 22 |
| Dar es salaam | 0.332 | 25 | 24 | 23 |
| Town West | 0.307 | 26 | 25 | 26 |

Note: Differences of three or more ranking places are shaded orange.

However, the importance of the real world, applied implications of measurement properties can be further illustrated by the identification of differences and poverty rankings for subgroups, of the population. We restrict our illustrative example to regional differences in poverty, and regional rankings by poverty in Tanzania. A fuller set of results is available from the authors. Table 3 shows the regional rankings for Tanzania, where there are 26 subnational regions/provinces. We calculate the headcount mean score for each index at this level and then compare all three indices using the sum-count index as the baseline. Poverty rankings that differ by *three places or more* from the sum-count index are highlighted in orange. The AF index produces three regions with a ranking difference of three or more places, while the CC index produces higher levels of ranking difference – 13 or one-half of all regions. These results suggest that the use of multidimensional poverty indices to regionally assess needs and allocate funds based on multidimensional poverty levels would potentially face huge uncertainty, especially when compared to the simple sum-count approach.

## 4. Findings and Conclusions

### 4.1 Findings

We have considered the performance of two methodologies for multidimensional poverty counting indices and compared them to a simpler sum count as a benchmark. We have done so with three main questions in mind for monitoring SDG poverty goals and targets.

How do the indices compare in their cardinal and scalar properties? This is at heart a rather academic question, but it has huge consequences for poverty measurement and thus to applied target and policy monitoring. We found that the AF index, using ten indicators, produced distributions that were normal but more granular. The number of increments in the scale depends on the differential weights but would be a minimum of ten. A CC index is a lot less granular because dimensions (categories) and not indicators are summed/counted but would always be less than the underlying number of indicators (10 in our examples). This has real repercussions for how poverty is interpreted because the underlying arithmetic link to the indicators of each deprivation is different between indices. A counting of categories (effectively categorical headings under which deprivations are aggregated) produces a numerical total that is more ordinal than categorical. This *can* be arithmetically analysed, but treating the mean or other summary statistic as if it came from a cardinal number raises some fundamental issues about how different *types* of variables were used for summing different underlying indicators in to very different categories. Multidimensional poverty indices obviously count different deprivations and thus require careful interpretation of the underlying differing natures of deprivation, for instance, of 'stock' and 'flow' deprivations, and of the different levels of deprivation. But the CC index adds another layer of uncertainty because it is

counting the categories/collections of such different indicators. We have moved from interpreting the sum of 'chalk and cheese' to interpreting the sum of their higher level categorization: of 'drawing media and dairy produce'. Using arithmetic summaries is thus beset with uncertainty and inexactness – what arithmetic meaning is there for the mean average of a count of categories of underlying objects? On the other hand, the AF methodology sums indicator weights and the index score are resultantly more cardinal in nature.

But for both indices it is important to note that a score may not reflect more or less underlying deprivation: it is possible for two children to differ in index scores for the same number of deprivations across both indices. However, good practice in MPI reporting often contains censored headcounts for comparison (see Alkire et al. 2017, for example).

AF index scores can always be decomposed back to indicator prevalence, but CC index scores cannot because dimensions are not derived by arithmetic sums but by Boolean aggregation. This means that an indicator in any dimension may or may not count, depending on how many other indicators it is in union with. Some have attempted to decompose CC indices by their categorical dimensions, using the same FGT decompositions established for the AF index. However, these categorical dimensions are difficult to arithmetically attribute to actual deprivation prevalence, and the alternative of decomposing by indicator is impossible.

Our laboratory testing also showed that CC indices tended to saturate easily. This means that arithmetic changes to the sum of dimensions are probably not consistent as the index changes to reflect the higher prevalence of deprivation and/or correlation between indicators of deprivation.

These properties lead us to consider the second question: How do the indices set robust baselines? Our analysis confirmed the theoretical literature's findings that a union approach produces an exaggeration of poverty in the CC approach: mean scores were higher by a factor of around 50% across both laboratory and real survey data examples. We do not suggest that the count of dimensions is not accurate, but that it skews the underlying prevalence of multiple deprivations upwards. We saw that no child was deprived in all ten indicators in the three countries, but that there were always children who were deprived in each dimension or category. Perhaps, the term reliability is more useful than robustness, but conclusions from this finding are of concern for applied policy measurement at the national level if poverty reduction planning is to be taken forward. In essence, the CC method is asking policymakers to assess the situation as worse than it actually is.

Finally, our third question: How do the CC and AF indices assess if poverty is changing over time to meet SDG targets? We found big differences between the indices in monotonicity: capturing change from underlying changes in indicator prevalence. Weights mattered and produced different levels of change

according to the assigned indicator weight in the AF index, but we also saw surprisingly different implicit dimension weights in what were normatively assigned equal weights in the CC index, reflecting the combination of household-level indicators and union properties. But differential weighting in the AF index was always seen to be symmetric and consistent: levels of change consistently reflected the arithmetic values assigned to the indicator as prevalence rose or fell. This was not so for the CC index, where the underlying logic of a Boolean union approach produced a range of cumulative effects. First, the same arithmetic property of exaggeration discussed above produces an over-representation of the likelihood of a move from zero to one when compared to a move from one to zero, especially in dimensions that have more than one indicator, which are the majority of dimensions in observed practice. Second, that property of asymmetry was mediated by saturation, making non-consistent asymmetry an axiomatic property of the index. Third, correlation matters hugely for indicators held in union for a CC index – a specific measurement property above and beyond the issue of overall correlation between indicators for all indices. Correlation within dimension leads to inconsistent changes to overall index score from changes in indicator prevalence.

How are these indices affected by the data properties of household clustering and age-specific censoring? Both increased the relative skewness of the CC index – increasing the probability of saturation, exaggeration, and non-monotonicity.

### 4.2 Discussion

What do our findings mean for child poverty measurement for the SDGs? Our findings suggest there are real appreciable differences in the ability of indices to capture levels of and changes to multidimensional child poverty. The important thing is thus to make such differences transparent to statisticians and policymakers in the countries that are considering them or that have put them in place. There is no such thing as a perfect poverty measure, but, given that the adoption of poverty goals and approaches is one for national actors, the differences in methods and their consequences for poverty measurement and setting poverty reduction targets must be made clear. There is a choice, and countries are free to make that choice, but when doing so they should be clear about the consequences of choosing a CC index over a AF index or vice versa.

Our approach suggests that when making the choice or when reporting poverty prevalence, the simple sum-count version of indices should be used as a simple sensitivity and robustness check – especially when important decisions have to be made about committing resources or assessing if policy is working or not.

We see our results as a first and preliminary step. There is still work to do to assess the optimal form of multidimensional poverty measurement. For instance, there needs to be a clear assessment of the issue of population weights, which is required when adjusting for differences due to age-specific censoring. The

issue of population reweighting deserves a paper of its own as it is too complex to cover here in any depth. But one early finding in the laboratory does suggest that differential indicator weights, as per the AF method, could be used to counter some of the effects of household clustering. This needs to be considered further and would mean a departure from the practice of normatively assigning weights to indicators.

But some future solutions to household clustering and age-specific censoring lie in better data. MICS and DHS programs are not designed to create multidimensional indices, but SDG targets now exist for larger age ranges of children and for more individual-level targets. This could eventually lead to the creation of suites of indicators that could create dimensions across all ages of children – for instance, considering cognitive development and other measures of non-cognitive performance for pre-school aged children that could allow 'learning' or some other higher-level dimension to replace the crudely determined 'education' dimensions that already exist.

Finally, we must re-emphasize our acknowledgement that national preferences for methodological approaches to poverty measurement may be politically driven in part. Our analysis has emphasized empirical measurement principles, but an alternative preference for counting rights or categories of poverty should also be acknowledged and respected if it drives political choices of method. As a final point we would like to suggest that it is possible to have one's political cake and eat it too when using robust empirical measurement approaches – there is always the ability to decompose and attribute multi-dimensional poverty to headings that capture the Convention on the Rights of a Child or other human rights headings. The AF methodology would allow this through decomposition rather than in its formulation. On the other hand, the CC approach puts rights first and foremost in the computation of the index, but at the considerable cost of monotonicity, exaggeration, and other measurement problems. The trade-off is not a zero-sum game, and the good news is that it is possible to have a rights-compliant index from DHS and MICS surveys that answers all of our three questions for SDG monitoring.

## 5. References


Alkire, S. (2014). 'Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index', *World Development*, vol. 59, pp. 251–274.

Alkire, S. and Foster, J. (2011). 'Counting and multidimensional poverty measurement', *Journal of Public Economics*, vols. 95(7–8), pp. 476–87.

Alkire S., Foster J., Seth S., Santos M-E., Roche J-M. and Ballón, P. (eds.) (2015). *Multidimensional Poverty Measurement and Analysis*, Oxford: Oxford University Press.

Alkire S., Dorji, L., Gyeltshen, S. and Minten, T. (2016). *Child Poverty in Bhutan: Insights from Multidimensional Child Poverty Index and Qualitative Interviews with Poor Children*, Thimphu, Bhutan: National Statistics Bureau.

Calderon, M-C. and Evans M. (2015). 'Multidimensional child poverty: How different approaches compare and potentially combine for practical and optimal policy engagement', Paper to Eastern Economics Association Conference, New York.

CEPAL (2013). *Panorama Social 2013*, Santiago de Chile: CEPAL.

CEPAL/UNICEF (2010). 'Pobreza infantil en América Latina y el Caribe', Fondo de las Naciones Unidas para la Infancia, Comisión Económica para América Latina y el Caribe, Publicación de las Naciones Unidas. Available *here*.

Chakravarty, S. and D'Ambrosio, C. (2006). 'The measurement of social exclusion', *Review of Income and Wealth*, vol. 52(3), pp. 377–398.

Commission on Global Poverty. (2016). *Monitoring Global Poverty: Report of the Commission on Global Poverty*, Washington DC: International Bank for Reconstruction and Development/The World Bank.

Dotter, C. and Klasen, S. (2014). *The Multidimensional Poverty Index: Achievements, Conceptual and Empirical Issues*, New York: UNDP. Available *here*.

Eurostat (2015). *Being Young in Europe Today: Living Conditions*, Luxembourg: Eurostat.

Foster, J., Greer, J., and Thorbecke, E. (1984). 'A class of decomposable poverty measures', *Econometrica*, vol. 52, pp. 761–776.

Gordon, D., Nandy, S., Pantazis, C., Permberton, S. and Townsend, P (2003). *Child Poverty in the Developing World*, Bristol: Policy Press.

Hancioglu, A. and Arnold, F. (2013). *Coverage in MNCH: Tracking Progress in Health for Women and Children Using DHS and MICS Household Surveys*, vol. 10(5), p. e1001391.

Hjelm, L., Ferrone, L., Handa, S. and Chzhen, Y. (2016). 'Comparing approaches to the measurement of multidimensional child poverty', Innocenti Working Paper 2016-29, UNICEF Office of Research, Florence.

Kovasevic, M. and Calderon, C. (2014). 'UNDP's Multidimensional Poverty Index: 2014 specifications', UNDP Occasional Paper, New York: UNDP HDRO. Available *here*.

Klasen, S. and Lahoti, R. (2016). 'How serious is the neglect of intra-household inequality in multidimensional poverty indices?' Courant Center Discussion Paper 200, Georg-August University, Göttingen.

MOLISA (2016). *Multidimensional Poverty in Vietnam*, Hanoi: Ministry of Labour, Invalids and Social Affairs.

Ravallion, M. (2011). 'On multidimensional indices of poverty', *The Journal of Economic Inequality*, vol. 9(2), pp. 235–248.

Rippin, N. (2010) 'Poverty severity in a multidimensional framework: The issue of inequality between dimensions', Courant Center Discussion Paper 47, Georg-August University, Göttingen.

Roche, J-M. (2013). 'Monitoring progress in child poverty reduction: Methodological insights and illustration to the case study of Bangladesh', *Journal of Social Indicators Research*, vol. 112(2), pp. 363–390.

UNDP. (2010). *The Real Wealth of Nations: Pathways to Human Development, Human Development Report 2010*, New York: United Nations Development Programme.