Multi-dimensional Poverty Indices for Children from Household Surveys: Lessons and ways forward[1]

Martin C Evans
Attila Hancioglu
Akmal Abdurazakov

Draft Working Paper Version
Not for quotation – circulate with care

Abstract: *This paper considers approaches used to measure child multi-dimensional poverty (MDP) in the developing world: the Alkire-Foster method and the 'CC' method as exemplified by UNICEF poverty indices based on methodologies by Gordon et al and De Neubourg et al. Discussion begins with survey micro-data extensively used for these indices, the Multiple Indicator Cluster Surveys (MICS) and the Demographic and Health Surveys (DHS), and the resulting data constraints on indices for measurement and coverage of MDP for children. Two important constraints are identified as affecting measurement of MDP across both indices: a) the inclusion of both household level and individual level indicators, b) the age-specificity of individual indicators for children and representation in survey data. Analysis considers the underlying differences between the two methodologies in two stages. First, using Monte Carlo simulations of hypothetical data we consider the differences in measurement properties that arise from axiomatic construction of indices, and the effects that 'household and individual' mixed level data and 'age specificity' have on such axiomatic properties. Second, we use harmonized DHS data from three countries to examine how those axiomatic differences in measurement properties affect MDP prevalence within and across countries, and the ability of indices to monitor changes in MDP prevalence. The paper concludes by considering the findings from the analysis and how they could be taken forward in the future collection and analysis of survey data for estimating MDP for children in Sustainable Development Goals targets and indicators, with particular reference to the MICS survey programme.*

Keywords: *Child poverty; MICS; DHS*

[1] <Acknowledgements here >.

Introduction

The inclusion of multi-dimensional poverty in the Sustainable Development Goals (SDGs) in 2015 was a long-awaited recognition of its importance and relevance, but has raised new, potentially exacting, requirement on measurement. The SDGs also prioritise children within poverty measurement making child multi-dimensional poverty a key area for SDG goals and targets. While children's multi-dimensional poverty indices may have been primarily constructed for advocacy purposes prior to the SDGs, they now have to perform as poverty measurement tools and thus have clear cardinal and scalar properties, to set robust baselines, and be able to capture changes in poverty over time to monitor SDG targets. This paper considers the underlying methodologies of approaches that are in place to meet this challenge.

Measurement of multi-dimensional poverty has grown rapidly since UNICEF's ground-breaking 2003 report on global multi-dimensional poverty for children (Gordon et al 2003). Other methods appeared, most notably the adoption by Human Development Report of the Multidimensional Poverty Index (MPI) using Alkire-Foster methodology in 2010 (Alkire & Foster 2011) for the Human Development Report (UNDP 2010). The A-F methodology (but not MPI) has now also been adopted by the World Bank implementing an index to meet the recommendation of the 'Atkinson' report (Commission on Global Poverty 2017). In Europe, Eurostat have developed multiple material deprivation measures that also consider non-monetary measures of poverty in a multi-dimensional 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, on characteristic of the literature is that multi-dimensional child poverty has been dominated by the latter approaches, and thus has mostly avoided the technical scrutiny of econometricians who dominate the former.

Our paper proceeds as follows. The remainder of this introductory section outlines the household survey data used for the indices then the multi-dimensional index methodologies that we compare. The analytical part of the paper follows in two parts: Part 1 considers 'laboratory' tests of the main indices in comparison to a simple 'sum-count' index as a benchmark; Part 2 considers the indices as implemented in actual household survey data in three countries to assess how far the 'laboratory' findings are present in real data. The paper then reviews its findings and makes some hesitant conclusions about both methodology and data.

Household Survey Data

We concentrate on the main sources of survey data that are currently feeding into large scale multi-dimensional child poverty measurement: USAID-supported Demographic and Health Surveys (DHS) and the UNICEF-supported Multiple Indicator Cluster Surveys (MICS). Current indices are constructed on data from these surveys undertaken before 2015, and thus not reflecting the SDG agenda. Both the MICS and DHS programmes were initiated prior to the Millennium Development

Goals, in 1995 and 1984, respectively (Hancioglu and Arnold, 2013). However, both programmes evolved to include all relevant MDG indicators in time, leading to the reporting of many MDG indicators based on results of MICS and DHS surveys for the majority of low and middle income countries. IN terms of content and household survey methodology, the two programmes display a large amount of similarities, and collaborate closely and work through interagency processes in an effort to harmonize survey tools and ensure comparability to the extent possible. However, there are also key differences between the two programmes, some of which are obvious differences and many subtle. In most cases, comparability of data from these two survey programmes is not compromised by methodological differences, which means that analysts can use data from both surveys to track trends in key indicators, especially since many countries regularly conduct both surveys, usually with reasonable intervals.

Both survey programmes offer a large amount of data that is or can potentially be used for multi-dimensional poverty analysis. The DHS programme focuses on data on health and population trends, with emphasis on fertility, family planning, mortality, reproductive health, child health, gender-related issues such as domestic violence, HIV/AIDS, malaria, and nutrition. MICS surveys provide key information on mortality, health, nutrition, education, HIV/AIDS, and child protection for use in programme decision making, advocacy, and national and global reporting. Both surveys are implemented by government agencies – in the case of MICS, almost all MICS surveys are conducted and owned by National Statistics Offices. MICS and DHS survey programmes regularly update and modify the contents of their questionnaires and frequently lead methodological developments in measurement of indicators in household surveys. Currently, both programmes are close to completion of inclusion of all relevant SDG indicators – both covering more than 30 SDG indicators, mostly overlapping.

One of the key differences between the two survey programmes is the way that the child population is covered. The MICS programme has traditionally included a separate under-5 questionnaire, administered to mothers, or in the absence of mothers from the household list, to caretakers, which ensures that in the presence of significant orphanhood and fostering, all children are covered by the survey. MICS has recently added a separate 5-17 Children's Questionnaire, again administered to mothers and caretakers with the same principle of full coverage in mind. The DHS programme also targets to cover all children; however, DHS does not include a separate questionnaire for children and obtains much of the information on children from their biological mothers.

Two key aspects of MICS and DHS data influence the performance of multi-dimensional poverty indices in practice:

- Surveys collect data at two levels: *individual and household*. Surveys collect common information on household and community level services – such as water, sanitation, and on household level resources – such as assets, the material construction of the home and demographic make-up of the household. Data on health, education and other areas of child and maternal well-being are collected at individual level – either from adult respondents or directly from child level

observations – e.g. anthropometrics. This means that many indicators of child poverty are clustered at the household level – all children are poor under that indicator if it is fulfilled at the household level, while there will be observed variance between children within households for child level data. Such clustering has serious outcomes for measuring differences at the individual level and can severely limit interpretations of gender, birth order or other individual level difference when both levels of indicators are joined into the same index.

- Surveys collect data for *age-specific profiles 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. 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 in children when such censored data is joined to household level data in indices – differences from age-composition of individual children now reinforce difficulty in measuring individual level differences that may already be obscured by clustering in households. The issue of age or population specific data and its effect on multi-dimensional indices was discussed at length by Dotter and Klasen (2015) and led to revisions of UNDP's MPI (Kovasevic and Calderon 2016)

The DHS and MICS programmes were never set up to ensure that indicators are present in sufficient numbers for household and individual indicators, and certainly not with the intention of capturing multi-dimensional poverty from an individual perspective, in the way that child multi-dimensional poverty is now defined. Since both programmes were designed before the advent of multidimensional poverty analysis and were based on key indicators in the sectors of concern, limitations as described above are natural. However, both surveys have been extensively used for such multi-dimensional analysis, and a recent development is the inclusion of derived multi-dimensional poverty indicators in the list of indicators of the MICS programme, which means that (1) the survey reports will be regularly producing estimates of multi-dimensional poverty, (2) the programme is likely to align closely to current and future developments in multi-dimensional poverty analysis and methodology

Multidimensional Poverty Indices

We limit our review to 'counting indices' drawn from DHS and MICS survey data. In these indicators are arithmetically summed according to a range of different weighting and aggregation assumptions. In its simplest form, a set of indicators for a counting index can be the sum of each indicator expressed in binary form. Thus, an index from 10 indicators, in this simple form, is the 'sum count' of deprivation indicators each child has, from zero to 10.

The assignment of indicators into dimensions is where methodologies differ, and differences arise in both the allocation of weights to indicators and/or dimensions and the approach to assigning indicators to and within dimensions. We consider two approaches

- A-F methodology (hereinafter 'A-F') is an index formed from a sum of binary indicators with no axiomatic adoption of weighting protocol. In applied

practice, indicator level weights are determined according the assignment of dimensions that classify indicators. Using such an approach, the sum of dimensions and indicators will be 1, but each indicator can be weighted independently or per share allotted to the dimension of which it is a part. In the most well know precedent, the global MPI, three dimensions were set to reflect the three dimensions of UNDP's Human Development Index (health, education and living standards) each having equal weight of 1/3. Education and health dimensions had two indicators each with resulting weights of 1/6; the 'Living Standards' dimension had 6 indicators, resulting in indicators weights of 1/18. But, while the MPI dominates discussion, *it should be considered a variant of A-F, not its essential representation.* For instance, Vietnam's Multidimensional Poverty measure, called MDP, (MOLISA 2016) has 5 dimensions and 10 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.

The setting of indicator cut-offs and poverty thresholds for the resulting index score leads to a poverty A-F headcount measure index and is accompanied by poverty metrics for 'intensity' and 'adjusted headcount' measures that allow a complete reconciliation of poverty measurement to the Foster, Greer Thorbecke (reference) standards for monetary poverty, and thus to intensity and 'poverty gaps and to most of the axiomatic requirement of poverty measurement established in the poverty literature (Alkire S et al 2015).

Children's multi-dimensional poverty can be captured by disaggregating household level MPI by age– as most recently done at the global level for the first time (Alkire et al 2017). But specific child level A-F measures also exist and, 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) and examples are currently underway in Vietnam, Maldives, Afghanistan, Malaysia, and other countries.

- Categorical Counting (hereinafter 'CC'). This term is ours and refers to a number of indices that use a normative 'rights based approach.[2] The crucial arithmetic differences to both A-F and the 'sum-count' approaches are four-fold

  - The *dimensions are counted* to produce the index score
  - Aggregation of indictors into dimensions uses a '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.
  - There is no necessity for consistent number of indicators per dimension. Some dimensions contain a single indicator (often 'sanitation' and 'water' dimensions in practice), while others can contain 2 or more indicators.
  - It is axiomatic that *each dimension has equal weight* arising from a normative rights-based assertion used in the approach

[2] Another term could be 'dimensional counting' but with A-F decomposition often producing results by dimension, we consider our term less ambiguous.

These indices are of longer standing – starting with Gordon et al's 2003 global child poverty profile (op-cit) and subsequent UNICEF global poverty and disparities profiles in around 50 countries, and a specific version of that index for Latin America (CEPAL/UNCEF 2010, 2012). More recently, Multiple Overlapping Deprivation Analysis (MODA) was developed using a similar approach (De Neuburgh et al 2013)

The Analysis

There is a small literature that directly compares these indices in practice. For instance, a comparison of individual level MODA indices compared to MPI household level indices for the same country (ies) (Calderon & Evans 2015, OoR paper). While these studies find differences in the level and composition of poverty, such differences can be difficult to interpret if it is not clear how they arise: from the underlying methodology, or from the construction in survey data (through use of different indicators, difference in the construction of similar indicators, or from underlying data cleaning work such as trimming of data for outliers, etc.). The MPI's early years was characterized by criticism of multi-dimensional assumptions and measurement outcomes from poverty economists established in the monetary approach (Ravallion 2011 and others). Over time a much larger analytical literature has arisen on the A-F approach and its comparison to statistical and econometric measurement practice of poverty in general. For instance, tests of robustness and sensitivity for MPI have been undertaken (Alkire 2014, for instance), 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 A-F and CC approaches perform in terms of three clear questions:

- How do they differ in cardinal and scalar properties?
- How do they set robust baselines?
- How do they asses if poverty is changing over time to meet SDG targets?

Our approach is to consider the properties that arise from axiomatic measurement principles in the first instance and then to assess how such identified properties affect performance in actual household survey data. We do not start with an algebraic proof for several reasons. First, we want to reach a wide audience that is more 'practice based' than the readership of highly technical econometric, mathematical and statistical journals. Second, algebra can sometimes be a 'black box' that hides uncertainty and assumptions. When considering counting indices, the Greek symbol sigma Σ may represent a cumulative sum of non-consistent components, and thus obscure differences between ordinal, categorical and cardinal interpretations. Third, our worked examples from 'laboratory data' benefit from and include the outcomes of a 'trial and error' process in some instances. Indeed, a lesson from our work is that demonstrating ex-ante theoretical measurement problems in the lab can identify unforeseen measurement properties. We end our analysis by moving from laboratory data to implement indices using real survey data and thus to 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 beyond simple descriptive comparisons of indices already in place and concentrate on 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 particular named indices but instead concentrate on the underlying measurement approaches and their assumptions. Our comparisons and review are of A-F, CC methodologies and generalizable properties of them. We recognize the investments that have gone into named indices but our analysis is of A-F and CC approaches, and not of the indices that spring from them: MODA, MPI, CEPAL/UNICEF etc.

Part 1:
Tests Using Laboratory Data

We construct laboratory data of 10,000 hypothetical observations with 10 non-specified binary indicators for each observation. We randomly (coin-flip) allocate ones and zero scores 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 in our sensitivity and monotonicity analysis. For the random data, we make comparisons between indices from 100 Monte Carlo trials of coin-flip random allocation to provide robust estimates at 99.9 percent level. This provides us a baseline distribution and result for a mean score of 0.5 across the 10 indicators against which to compare index performance.

We have three test versions of indices:

- A-F index. Our test index uses the same weights as those in the global MPI discussed earlier – reflecting noth 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.
- The CC index. Our test index uses weights and approaches from the applied literature (Gordon et al 2003 and De Neuburg et al 2013).
  - Five dimensions:
    - three dimensions are populated by 2 indicators in 'union approach'
    - one dimension has 3 indicators in union approach and
    - one dimension has just one indicator.
- The 'Sum-Count' index. This is for comparison and baselining purposes.
  - Ten indicators each with equal weight of 0.1.

Tests for Index Scales and Baselines

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

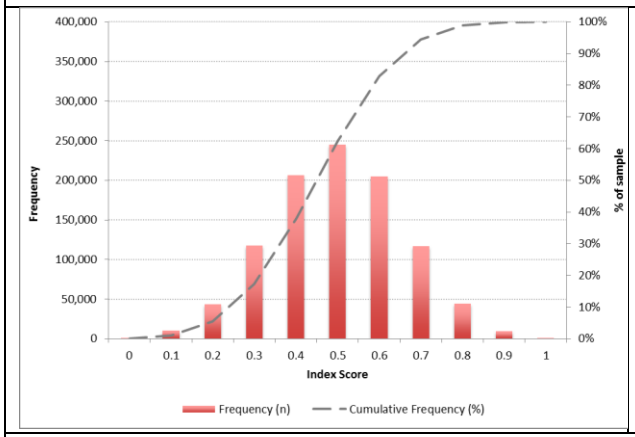
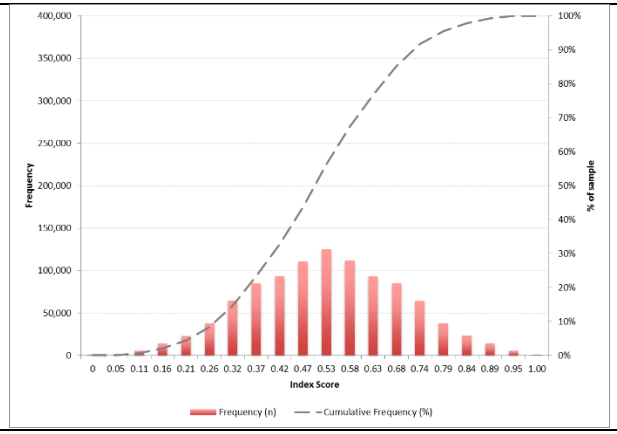
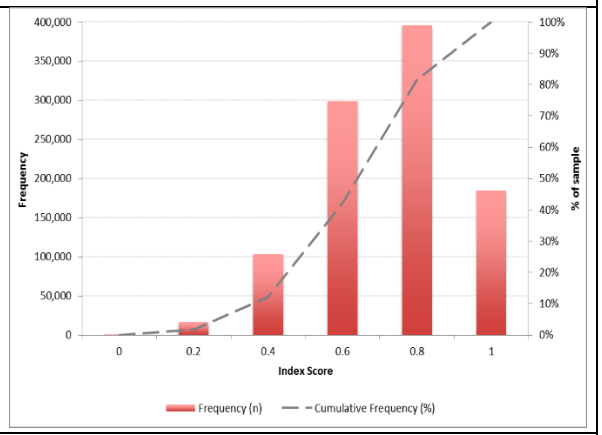
Figure 1b) shows the A-F index also results in a bell curve distribution, (but Shapiro-Wilk tests do *not confirm* it is a normal distribution) with 18 equal

increments of 0.566 (the value of the smallest weight) between 0 and 1. Having index weights smaller than 1/10 results in a more granular distribution compared to the 'sum count' benchmark but with consistent mean and median at 0.5 and a standard deviation of 0.18.

Figure 1.c) shows the results for the CC index, which stand in stark contrast to both the benchmark and A-F specifications for the same data and underlying distribution of deprivation scores. The first thing to note is the far less granular distribution, as counting 'dimensions' rather than summing indicator scores gives only 5 increments of 0.2 between 0 and 1 on the scale. But most noticeably, the distribution for the CC approach is hugely skewed and has resulting higher index scores overall: 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' prevalence of multi-dimensional 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 identified that 'union approach' results in '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 verses 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 A-F 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*, and the finding that CC exaggerates poverty at all thresholds compared to the A-F and 'sum count' specification in any form. Confirmatory results can be obtained from the authors.

Figure 1 Baseline Results
Monte Carlo Simulations (100 trials)

‘Sum-Count’- Benchmark		**Alkire-Foster**		**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

Tests for Sensitivity and Monotonicity

Our findings so far on the shape and properties of the different distributions formed by indices also raise findings that are relevant to our second 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 difference between sub-groups and for tracking change over time.

Figure 4 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 – symmetry. Figure 4 shows that the Sum-Count index, with every 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 A-F index has a similar symmetrical profile but produces larger changes in index scores for the same incremental change in indicator prevalence: from 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.

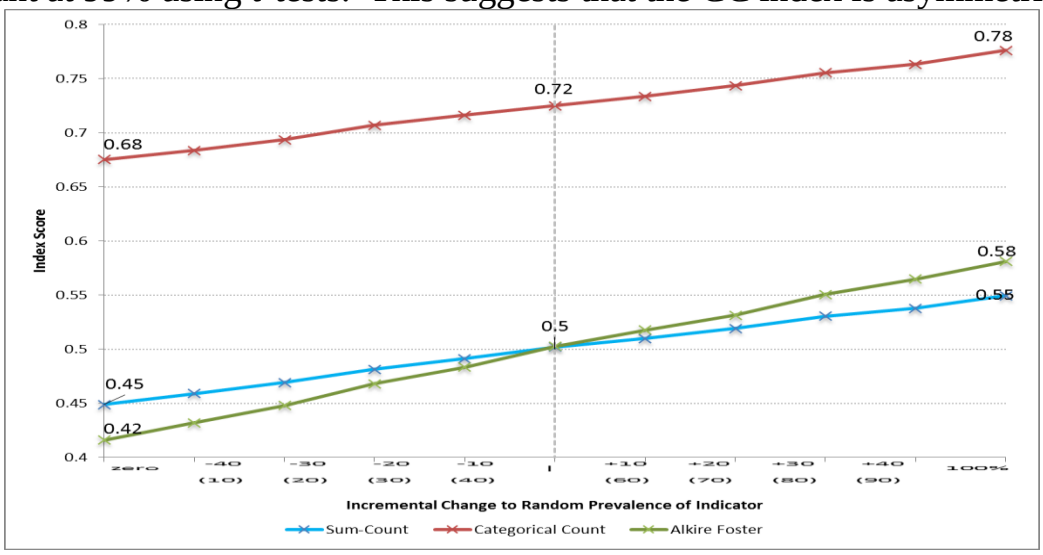


Figure 2. Changes in Index Scores from Incremental Change to Any Indicator
(100 Monte Carlo Trials)

Figure 3 shows the influence of differential weights in the A-F index. The level of incremental change differs: large change is the result of higher indicator weight, but changes similarly symmetric across incremental change to indicators of differential weight.

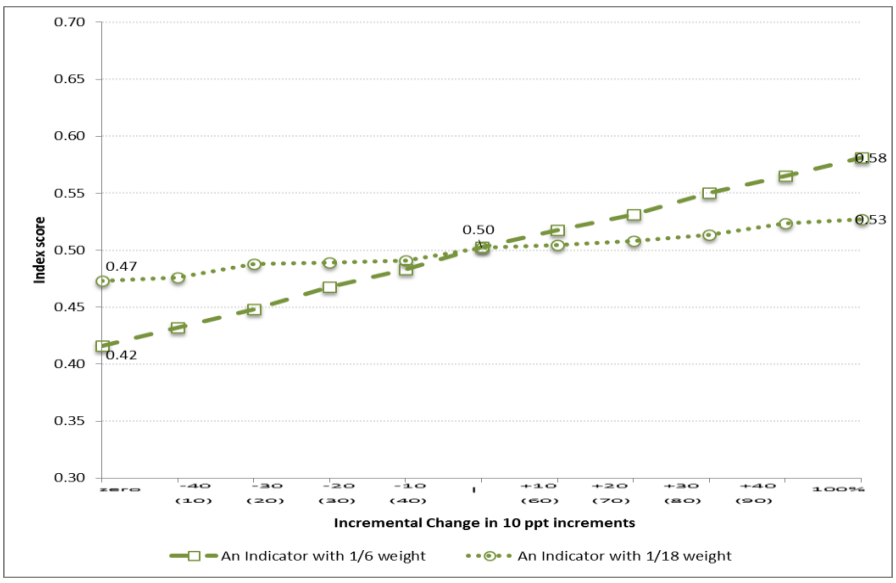


Figure 3. A-F: Incremental Change in Indicator by Assigned Weight
(100 Monte Carlo trials)

When we attempt to replicated 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 laboratory data sets all indicator prevalence to 0.5, with random, independent (non-correlated), indicators. This effects for the CC index because we find saturation effects, which additionally prevent the index from increasing overall scores from underlying increased prevalence.

Figure 4 shows the effects of saturation and the underlying patterns of asymmetry that are seen in the CC index. same random assigned and non-correlated lab data used at different levels of randomly assigned prevalence in order to account for ‘saturation’ from our initial randomization that gave 0.72 as the mean starting point. The results show clear asymmetry but with asymmetric attributes that differ both by levels of saturation and by the ‘union assumption’ for the indicator of 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 from the indicators according to their ‘union’ with other indicators. As expected, and by construction, the indicator that has no other indicators in union (‘single union’), has clearest monotonic characteristics across all levels of saturation. However, the indicators in union with one or two other indicators clearly show differential slopes and very much small levels of incremental change over all the ranges of incremental change of indicator, and become ‘flat’ at high levels of concentration – where little if any change to the overall score from incremental changes in indicator prevalence. These characteristics of asymmetry operate across the implied differences in level of change that occur from the indicator’s implied ‘weight’.

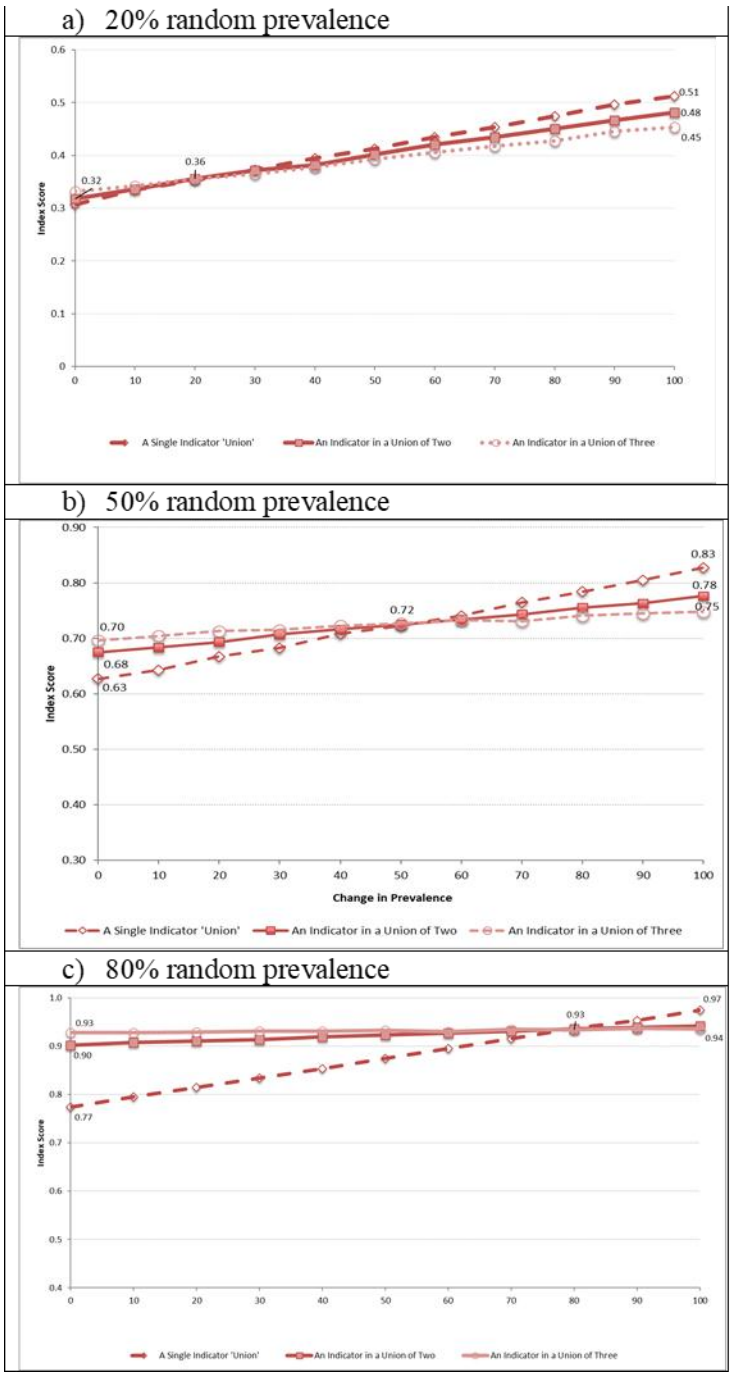


Figure 4. CC: Incremental Change in Indicator by Union Assumption with 20%, 50% and 80% random prevalence
(100 Monte Carlo Trials)

Testing the Effects of Correlation

So far we have relied on independent randomly assigned prevalence across our 10 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, A-F 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.

obs	A	B		Union A & B	C		Union A&C
1			negative correlation			positive correlation	
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%		60%	20%		50%

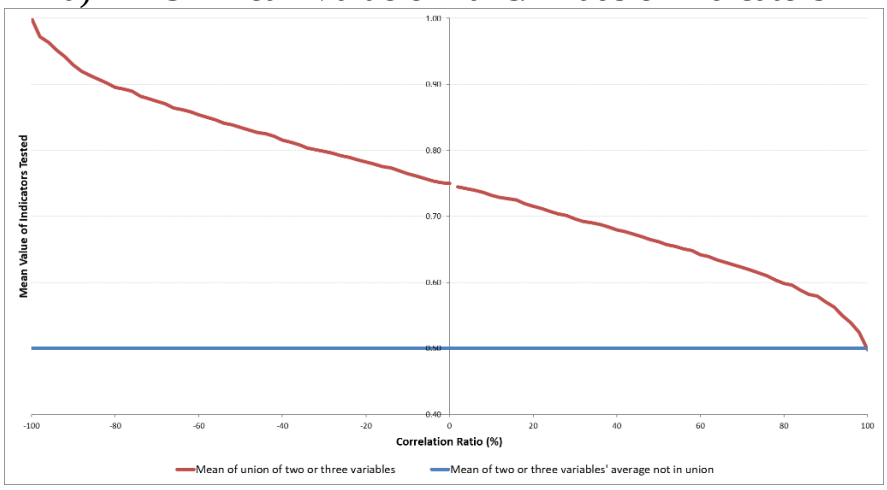
Figure 5. Differential outcomes of Negative and Positive Correlation in Union Aggregation

Figure 6 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 due to B being populated in observations that are not populated by 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 to observations that are also deprived in 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 to indicator prevalence will not produce increases to overall index score due to positive and negative correlation differences.

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 general 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 100fixed intervals of the correlation ratio, representing 2 percentage point increments between negative 100 and positive 100 percent. 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 5 show two tests that assess how the level of correlation affects indices. Figure 5a shows results for pairs and triads of indicators, and Figure 5b shows the results of correlation of pairs on the overall index scores. We set the underlying average indicator prevalence for the data to 50% prevalence for both these tests.

a) On Mean Value of Pairs/Triads of Indicators



b) On Index Scores from Correlation of Indicator Pairs

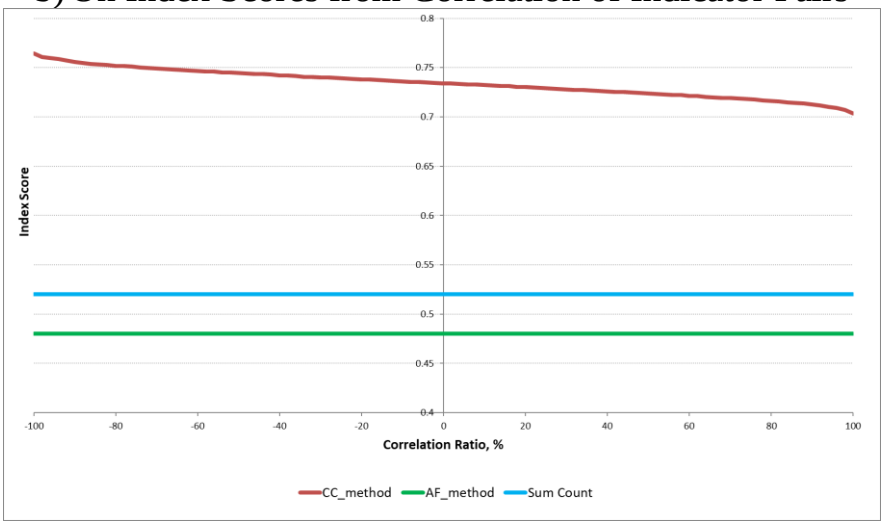


Figure 5. The Effect of Correlation on Multi-Dimensional Counting Indices

Figure 5a 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 aggregated using the union approach – either as pairs (twofold union) and 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 5b 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 as before at 50% prevalence. We thus test to see how the indices react to positive correlation between 2 indicators out of all ten indicators. Figure 5b demonstrates how the CC index score is influenced non-linearly by the level of correlation between those two indicators. Neither A-F nor Sum Count indices are influenced by differing correlation levels between that identical pair of indicators.

Out 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 6 shows the results. Correlation is seen to make 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, for the same level or correlation for the same changes in prevalence for such correlated indicators.

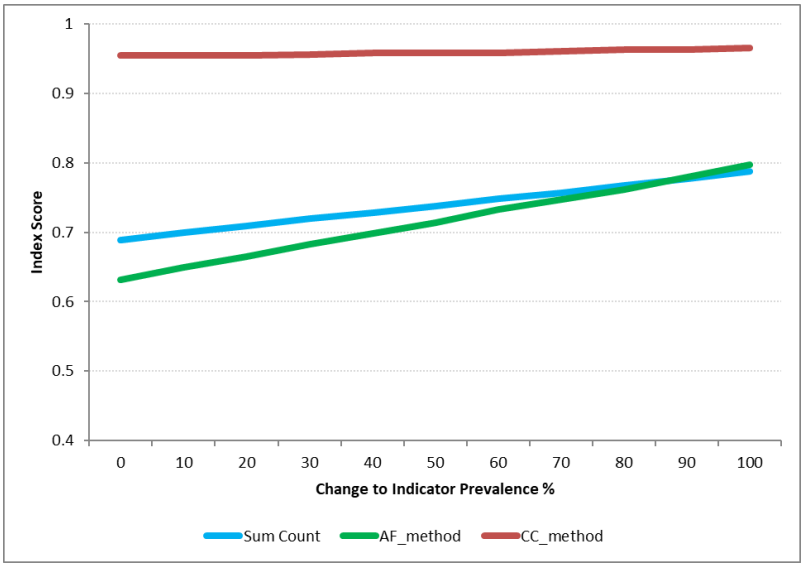


Figure 6. Changing Indicator Prevalence to High Correlation

Household Clustering, Age-Specific Censoring

We return to our original laboratory dataset of 10,000 observations and 10 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 child level indices,

- That indicators can be clustered at the household level while others are at the individual level
- That individual level indicators are age specific and any observation (individual) not in that age range is missing and thus censored for that indicator

We take these two issues in term.

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:

- We set the average number of observations per pseudo household as 3,
- We allow up to a maximum of 7 observations to be allocated to a random household.
- To this new distribution, 2 indicators were assigned to be household level indicators, the remaining 8 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 our findings on index reaction to clustering was unchanged in nature.

The results are shown in Figure 7. 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 desicussed earlier. On the other hand, the Sum Count and A-F indices are leaner functions, and the inclusion household level observations produces an upward shift of 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 A-F, a reflection that, on average, lower weighted indicators were affected. This raises the possibility that A-F type differential weights could be used to address household clustering effects on individual level indices. We return to this point later in discussion.

But our test on the impact of household clustering of 2 out of 10 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).

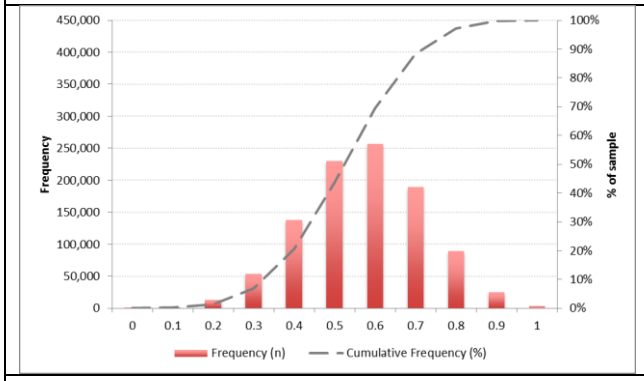
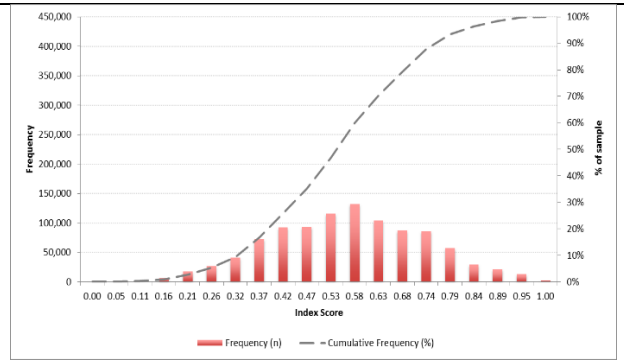
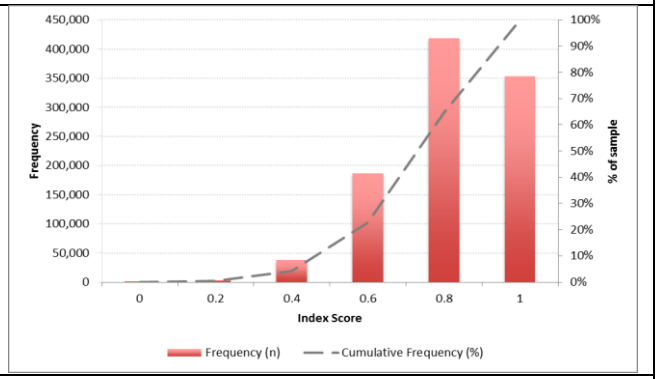
Sum-Count Benchmark		**A-F[3]**		**CC[4]**	
					
Summary Statistical Properties & Differences from Baseline					
Revised Statistic	Difference from Baseline	Revised Statistic	Difference from Baseline	Revised Statistic	Difference from Baseline
Mean 0.57	+0.07 **	*Mean* 0.54	+0.04 **	*Mean* 0.82	+0.09 **
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	

Figure 7 Revised results allowing for Household Clustering (2/10 indicators at household level)

Monte Carlo Simulations (100 trials)

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

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

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

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

Table 1

Household Clustering at Higher Margin (6/10 indicators household level)

	Sum-Count	A-F[5]	CC [6]
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

Table 1 gives an indication of the additional difference that would from such higher levels of household clustering – with 6 out of 10 indicatorts at the household level. The CC index's saturation is now clear, with mean at 0.93 and median at 1. These results suggest a clear caveat for using indices of this type in contexts with high levels of deprivation and high proportion of indicators at household level. On the other hand, A-F 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.

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 0-4 year old population in the three countries (UNDESA 2015) that we consider later in Part 2. To do this we replace all postive 1 values to zero for these 2,900 cases. While in reality, these values would be 'missing', we chose to replace with zeros to avoid the need to additionally re-weight to adjust to underlying population differences. However, in applied use of indices, population reweighting would require more consideration and we discuss this issue later below.

Figure 8 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 7. The results show 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.

[5] Dimension1(I;HH), Dimension 2(I;HH) and Dimension 3(I;I;HH; HH;HH;HH)
[6] Dimension 1(I;HH) Dimension 2(I;HH), Dimension 3(I;HH), Dimension 4(I; HH;HH) and Dimension 5(HH)

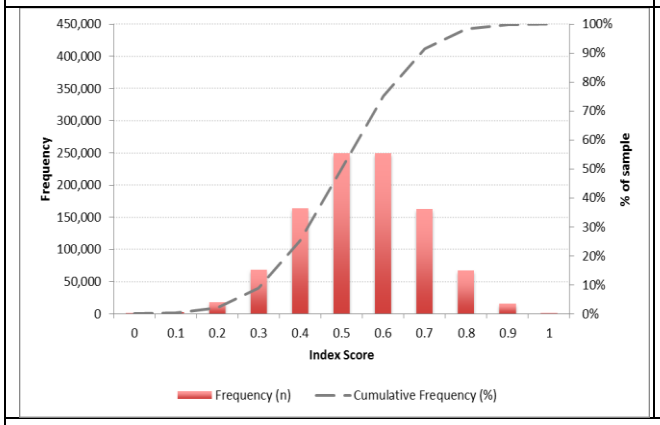
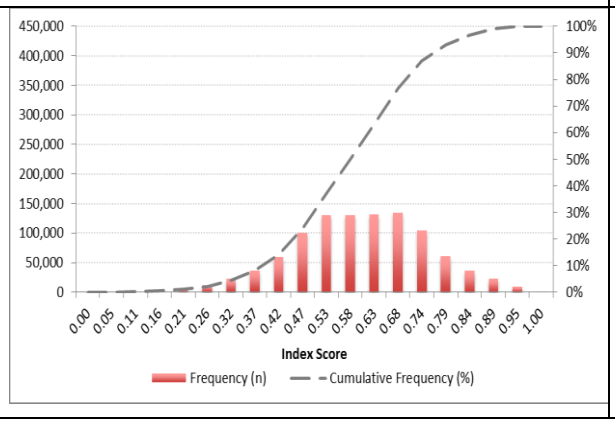
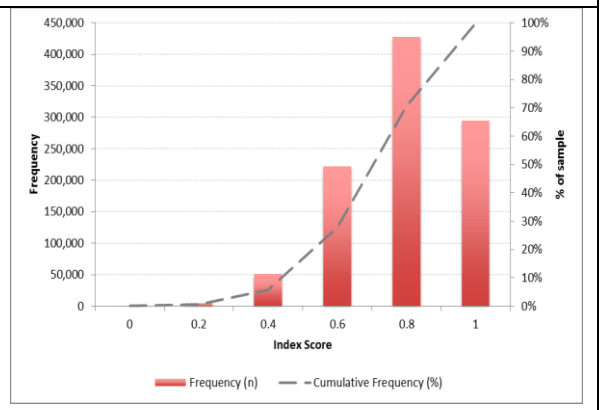
Multiple Deprivation- Benchmark		A-F		CC	
					
Summary Statistical Properties & Differences from Iteration shown in Figure 7					
Revised Statistic	Difference from Figure 5	Revised Statistic	Difference from Figure 5	Revised Statistic	Difference from Figure 5
Mean 0.55	-0.025 **	*Mean* 0.58	0.039 **	*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	

Figure 8. Results allowing for Household Clustering and Age-Specific Incidence of Deprivation
Monte Carlo Simulations (100 trials)

Notes: ** significant at 1% using two tailed t-test
The results from alternative allocation of household level indicators to dimensions are available from the authors but do not alter interpretation of results from this example

For A-F 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 increased in standard deviation. However, underlying skewness measures have risen to the highest level across all three simulations across Figure 1, 7 and 8. These results show the CC index reacts differently to changes in individual level prevalence compared to the Sum Count and A-F indices and this will further contribute to problems of monotonicity

We now turn to a final consideration of the potential effects of clustering and censoring by looking at 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 certain age. How does clustering and censoring affect the assessment of indicator correlation, and how can our indices minimize the effects of correlation bias.

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 changes from zero to missing
- Changes sign when household variables rewieghted
- Changes sign when both adustments 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

Figure 9. Correlation Tests and Clustered and Censored Indicators

In Figure 9 we visually demonstrate how an assessment of correlation is affected by clustering and censoring. We show hypothetical data (not randomly assigned) from the first 5 households of a hypothetical dataset with 10 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 9 instances shaded blue. The third matrix shows the results of re-weighting the values of clustered household level indicators, using a simple per-capita approach (for a household with 4 observations we 'reweight' the indicator value from 1 to 0.25 (¼), and so on). Again, this adjustment changes the sign of 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 highlighted orange. This test and designed to visually show the potential 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 of value of correlation ratios or on the robustness of changes in sign, but actual survey data of significant sample size would show similar changes to the size and sign of correlation ratios.

What effects will these changes in correlation 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 of resulting bias, as we have seen how different signs of correlation between indicators have the ability to alter monotonicity most in that index. But CC indices, due to their axiomatic normative adoption of 'rights based' equal weights, prevents changing index weighting assumptions to empirically reflect effects of clustering or censoring. The practice for these indices is not consistent, but the more recent MODA approach (de Neuburgh et al 2013) has produced different indices for age-specific sub-group of children unlike earlier CC approaches (Gordon et al 2003, UNICEF/CEPAL 2009). On the other hand, A-F approaches are more adaptable to empirically derived indicator weights to 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 global MPI (Klasen & Dotter 2014, Kovacevic and Calderon 2015).

Part 2: Indices using Household Survey Data

In the second part of our analysis we test some of the key findings from the laboratory work with real survey data. At this point it is crucial to restate that this work will NOT replicate actual indices that are in place. We test underlying methodologies not compare 'branded indices' of different forms such as MPI, MODA or others.

We use harmonized survey data prepared by Professor D. Gordon and his team at 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 DHS surveys in 2010 for Colombia, Bangladesh and Tanzania. To avoid some of the problems of age-specific censoring and in the interests of space and concision, we limit our indices to the population aged less than five years old. We take 10 indicators and construct three indices from those indicators:

- the 'sum-count' benchmark index.
- A-F: we have 3 dimensions (2, 3 and 5 indicators per dimension)
- CC: 5 dimensions (2, 3,1,1,3 indicators per dimension)

Table 2

Indicators, Dimensions and Weights for Test Indices

Categorical Counting: Composition of Dimensions					
Nutrition **(1/5)**	Infant feeding	Wasting			
Health **(1/5)**	DPT all	Unskilled birth attendant	Child mortality		
Water **(1/5)**	Drinking water				
Sanitation **(1/5)**	Toilet type				
Housing & Living Standards **(1/5)**	Overcrowding	Wealth low quintile	Info devices		
AF method: Composition of Dimensions					
Nutrition **(1/3)**	Infant feeding (1/6)	Wasting (1/6)			
Health **(1/3)**	DPT all (1/9)	Unskilled birth attendant (1/9)	Child mortality (1/9)		
Living Standards **(1/3)**	Drinking water (1/15)	Toilet type (1/15)	Overcrowding (1/15)	Wealth poorest quintile (1/15)	Lack of Information Devices (1/15)

Our indices are not designed to be relevant or to be inherently robust or meaningful in 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 multi-dimensional poverty for this test, merely in their comparative performance according to underlying measurement properties. Our choice of indicators is also dictated by the need NOT to approximate to an actual index in place. We have chosen some indicators that are used in multi-dimensional indicators, and others, such as 'lowest quintile of wealth index' that are not, and perhaps, never should be.

Figure 10 shows the key results for each of the three indices in each of the three countries and we limit reporting of results to simple comparison of means and maxima in order to establish whether the lab results of 'exaggeration' of poverty we saw earlier continue to be seen between CC and the other indices using actual survey data. The Sum Count and A-F indices all provide similar mean headcounts and similar maxima – using these combination of deprivations, the maximum score is 0.8 across all countries for the 'sum-count' index and 0.87 for A-F. However, the CC index gives consistently higher mean scores compared to the other indices – in the region of 50 per cent higher. Additionally, the maximum score for CC is always 1.00,

representing the outcome from counting ‘dimensions’, or headings of 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 verses A-F and the Sum Count’.

How do the indices perform when considering monotonicity? We show tests for two indicators: water and the presence of 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 CC as it is a single variable dimension, but has lower implicit weight in the other indices.
- Presence of an unskilled birth attendant is an individual level variable that has low implicit weight in CC as 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.

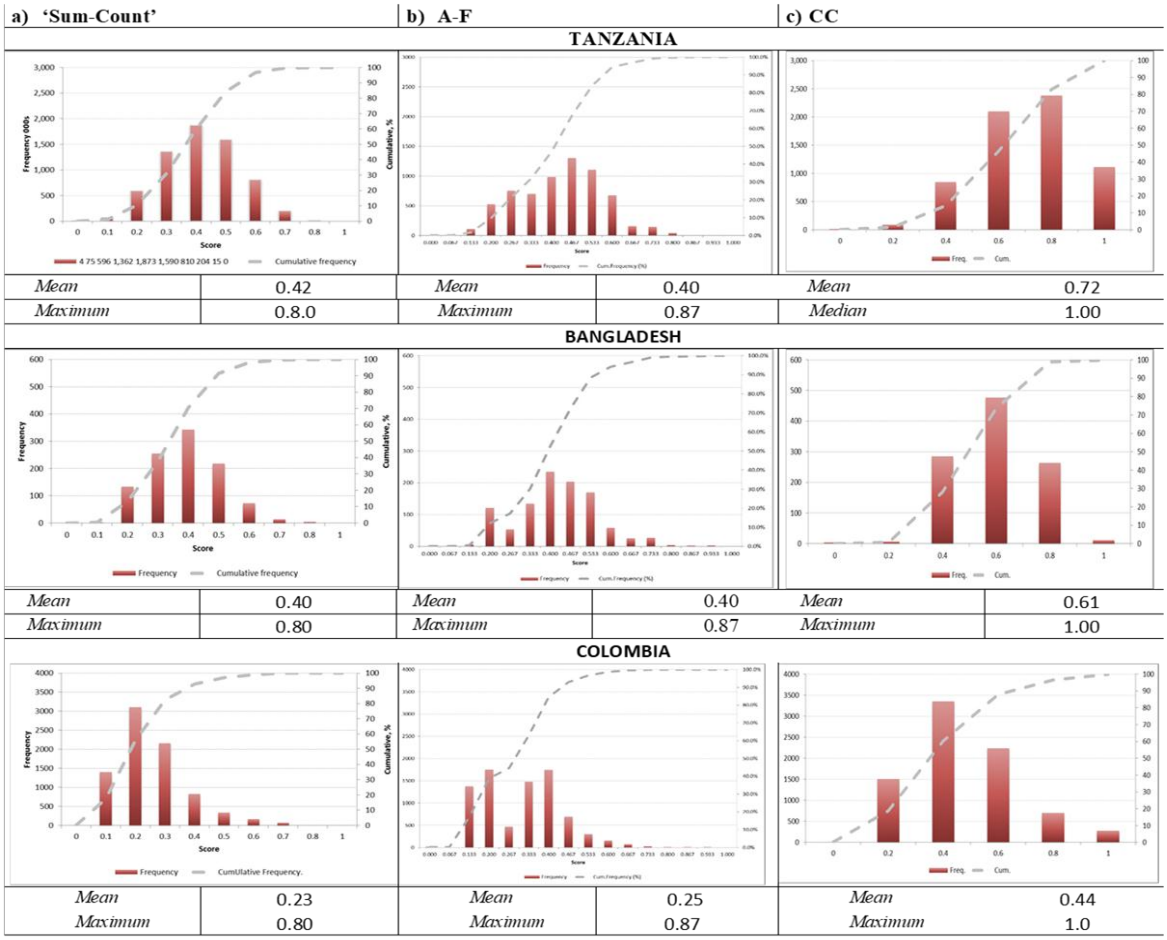


Figure 10. Headline Results for Indices in Three Counties: distributions, means and maximum values

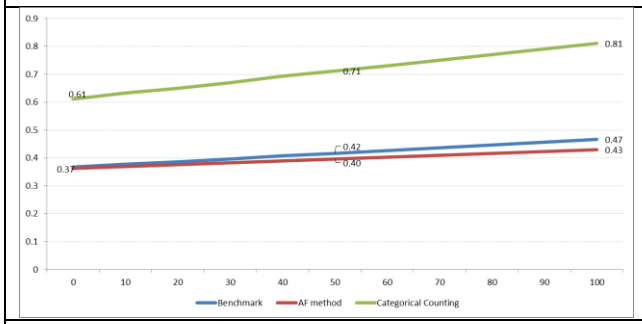
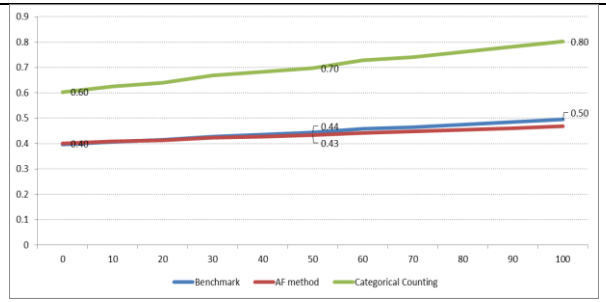
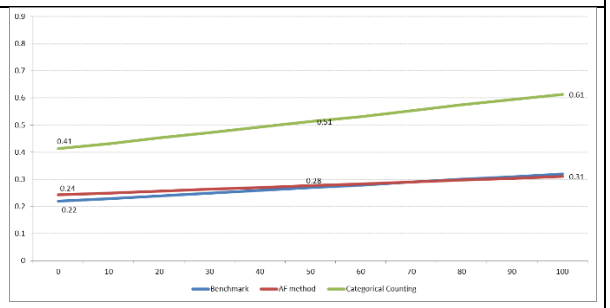
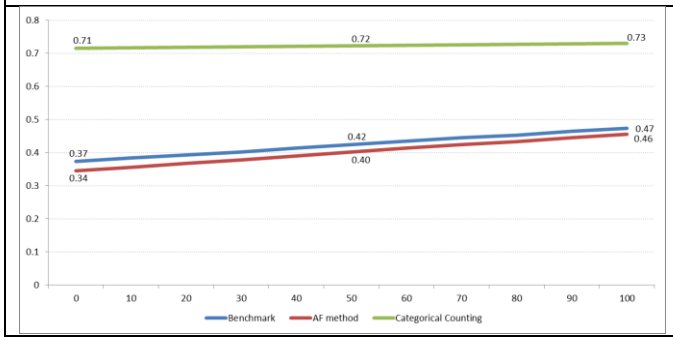
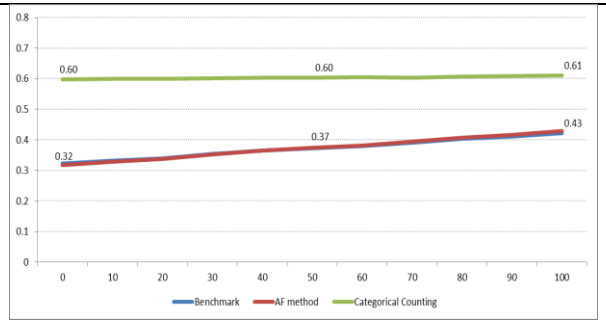
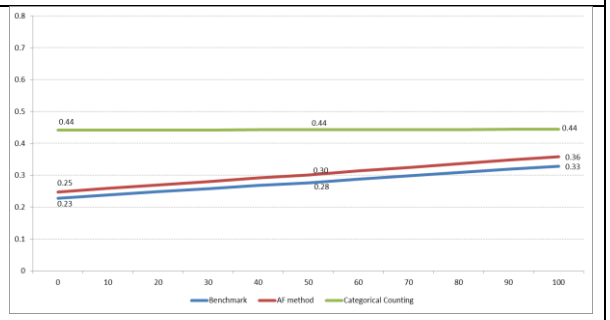
TANZANIA	BANGLADESH	COLOMBIA
Water		
		
Presence of Skilled Attendant at Birth		
		

Figure 11. Sensitivity Tests for Water and Presence of Skilled Birth Attendant Indicators

Figure 11 shows common results across indices and across countries in the changes to both indicators from incremental changes in prevalence. Incremental changes to prevalence of the water indicator has 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 A-F 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, discernable change to CC index – as shown by the 'flat line' in Figure 11, but 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. Investments in service provision could either see huge or no credit given in changing poverty indices using the CC method.

Table 3

Tanzania: Regional Ranking for Multi-dimensional Poverty Index Scores

Multi-dimensionally Poor Children aged under 5

		Rank Sum		
	Sum Count	Count	Rank A-F	Rank CC
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 nort	0.373	20	16	24
iringa	0.359	21	22	17
zanzibar sout	0.353	22	21	25
ruvuma	0.347	23	23	20
kilimanjaro	0.337	24	26	22
dar es salaan	0.332	25	24	23
town west	0.307	26	25	26

Difference in 3 Ranking Places Shaded

However, the importance of real world, applied implications of measurement properties can be further illustrated with real survey data and we consider a single further instance of potential importance: the identification of differences and poverty rankings for sub-groups, 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 sub-national 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. Where poverty rankings differ by *3 places or more* from the sum count index we highlight in orange. A-F produces three regions with ranking difference of 3 or more places, while CC produces higher levels of ranking difference, 13 or one-half of all regions. These results suggest that the use of MD poverty indices to regionally assess needs and allocate funds based on multi-dimensional poverty levels would potentially face huge uncertainty, especially when compared to the simple 'sum count' approach.

Findings and Conclusions

Findings

We have considered the performance of two methodologies to multi-dimensional 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 has huge consequences for poverty measurement and thus to applied target and policy monitoring. We found that using 10 indicators A-F 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 10. CC is a lot less granular because dimensions (categories) and not indicators are summed/counted but would always be less than 10 for that number of indicators. 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 headings under which deprivations are placed) produces a categorical ordinal variable which, of course, can be counted but interpreting the sum as a cardinal number needs a lot of care. A-F sums indicator weights and the index score is resultantly more cardinal in nature. But for both indices a score may not reflect more or less deprivation: it is possible for two children to differ in index scores for the same number of deprivations across both indices. However, it is noticeable that good practice in MPI reporting often contains 'censored headcounts' for comparison (see Alkire et al 2017 for example).

A-F index scores can always be decomposed back to indicator prevalence but CC cannot because dimensions are not derived by arithmetic sums but by Boolean aggregation: an indicator in any dimension may or may not count depending on how many other indicators it is in 'union' with. Practice in CC has established non-consistent aggregation between dimensions, making arithmetic attribution at the indicator level very problematic.

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

These properties lead us to consider the second question: *How do they set robust baselines?* Our analysis confirmed the theoretical literature's findings that 'union' approach produces 'exaggeration' affected the CC approach: mean scores were higher by a factor of around 50 per cent 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 deprivation upwards. We saw that no child was poor in every one of 10 deprivations in three countries, but that children we always found to be poor under every heading. Perhaps, the term 'reliability' is more useful than robustness, but conclusions from this finding are for applied policy measurement at the national level to take forward.

Finally, our third question, "*How do they asses if poverty is changing over time to meet SDG targets?"* We found big differences between the indices in capturing change from underlying changes in indicator prevalence. Weights mattered and produced different levels of change according to the assigned indicator weight in A-F, but we also saw surprisingly different 'implicit' dimension weights in what were normatively assigned 'equal weights' in CC index, reflecting the combination of household level indicators and 'union' properties. But differential weighting in A-F 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 CC index where the underlying logic of Boolean union approach produced a range of cumulative effects. First, the same arithmetic property of exaggeration as discussed above produces and 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 CC – 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 CC – increasing the probability of saturation, exaggeration and non-monotonicity.

Discussion

Our approach strengthens the case for using the simple 'sum-count' version of a set of indicators alongside indices when comparing them. Indeed, we would strongly suggest that this be a simple 'sensitivity and robustness' exercise when testing indices before they are adopted for measurement purposes.

We directly considered the impact of household clustering and age-specific censoring in the laboratory but not in the analysis of the three national datasets. One reason for this was insufficient space and time, and thus we have left some issues for

future work. But another constraint was tackling the issue of population weights, which would be required for adjusting differences from age-specific censoring. The issue of population reweighting deserves a paper on its own and was not collapsible to cover here in any depth. But one early finding in the laboratory does suggest that 'differential indicator weights', as per A-F, could be used to counter some of the effects of household clustering. This needs to be considered further, and would mean a departure from practice in which equal weighting was normatively assigned. The issue would be how far replacing 'equal weights' with empirical assumptions makes better child level indices at the expense of complexity and transparency and thus spoil the 'easy sell' to policy makers.

But the future to some of the solutions to household clustering and age-specific censoring is through better data. MICS and DHS programmes are not designed to make multi-dimensional 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. The example of ensuring no age-censoring in most of Bhutan's child MPI (Alkire et al 2016) is a clear pointer on how to bring together different indicators to cover all children of all ages consistently. This was a methodological solution to a measurement problem that was not overly constrained by fixed normative labels for dimensions, a clear indication of pragmatic ways forward. Other issues for survey data are indicators or material deprivation – in these or other surveys. Better individual age-related population weights in survey data is also a clear need for the future.

But finally, we must emphasise our acknowledgement that national preferences for methodological approaches to poverty measurement are at the heart of SDG poverty reduction. We have emphasized empirical measurement principles but an alternative preference for counting 'rights' or categories of poverty should also be acknowledged and respected. For poverty statisticians facing this choice, the need is to ensure that such preferences are accompanied by transparent knowledge of and acceptance of the outcomes of choosing a methodology. The empirical and measurement consequences of that choice are what we have tried, in part, to outline here. *But it is always the case that you can attribute multi-dimensional poverty to breaches of rights through a decomposition of an index rather than in its formulation*. Our findings suggest that the measurement of poverty through CC of rights does not allow the opposite to be true. Thus the trade-off is not binary but the good news is that it is possible to have a rights compliant index from DHS and MICS surveys that answers all our 3 questions for SDG monitoring.

Bibliography

1. Alkire S (2014) Measuring Acute Poverty in the Developing World: Robustness and Scope of the Multidimensional Poverty Index, *World Development* Vol 59, July 2014, Pages 251-274

2. Alkire S and Foster J (2011) Counting and Multidimensional Poverty Measurement, *Journal of Public Economics*, 95(7-8) 476-87

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

4. 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.

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

6. 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, https://www.unicef.org/honduras/Pobreza_infantil_America_Latina_Caribe_2010.pdf .

7. Chakravarty and D’Ambrosio (2006) The measurement of social exclusion, *Review of Income and Wealth* 52(3), 377-398

8. 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.

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

10. http://hdr.undp.org/en/content/multidimensional-poverty-index-achievements-conceptual-and-empirical-issues

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

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

13. Hancioglu A, Arnold F (2013) Coverage in MNCH: Tracking Progress in Health for Women and Children Using DHS and MICS Household Surveys. *PLoS Med* 10(5): e1001391. doi:10.1371/journal.pmed.1001391

14. Kovasevic M and Calderon C (2014) UNDP’s Multidimensional Poverty Index: 2014 Specifications, New York: UNDP HDRO http://hdr.undp.org/en/content/undp%E2%80%99s-multidimensional-poverty-index-2014-specifications

15. Klasen S and Lahoti R (2016) *How Serious is the Neglect of Intra-Household Inequality in Multi-Dimensional Poverty Indices?* Courant Center Discussion paper 200, Göttingen, Georg-August-Universität Göttingen.

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

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

18. Rippin N (2010) *Poverty Severity in a Multidimensional Framework: The Issue of Inequality between Dimensions*, Courant Center Discussion Paper 47, Göttingen, Georg-August-Universität Göttingen.

19. 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, Issue 2, pp 363-390

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