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

Draft for Conference Presentation: Panel 'Age and Gender-specific Poverty' at the International Statistical Institute's 2017 >

Martin C Evans
UNICEF, New York, USA – mcevans@unicef.org

Attila Hancioglu
UNICEF, New York, USA – ahancioglu@unicef.org

Akmal Abdurazakov, consultant, UNICEF

Abstract

This paper considers approaches used to measure child multi-dimensional poverty (MDP) in the developing world: the Alkire-Foster method and the 'categorical counting' 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] <Aknowledgements here >.

Introduction

This paper is presented in a panel of papers for the conference session: 'Age and Gender-specific Poverty' at the International Statistical Institute's 2017 conference in Marrakesh. This version of the paper focuses on responding to the issues raised by other papers in the session and omits discussion of the wider literature and other elements of a full academic paper. A fuller paper is available from the authors.

UNICEF was in the vanguard of international multi-dimensional poverty measurement with the 2003 report on global multi-dimensional poverty for children (Gordon et al 2003). But subsequently, the measurement of multi-dimensional poverty has grown rapidly. Other methods developed, most notably the adoption by UNDP of the MPI using Alkire Foster methodology in 2010 (Alkire & Foster 2011) for the Human Development Report (UNDP 2010). The Alkire Foster methodology has now also been adopted by the World Bank who are currently designing and implementing such an index following 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) but favour a simple 'multiple deprivation' count that does not involve adjusted weights or impose 'dimensionality'. These three approaches will be explored in consistent terms in this paper.

We face a measurement literature that has also expanded exponentially but across a huge gulf in 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. One result is that multi-dimensional child poverty has mostly avoided the technical scrutiny of econometricians, who have focused on non-specific age-based indices in general.

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. While previously children's indices may have been primarily constructed for advocacy purposes, they now have to perform as poverty measurement tools and thus have clear cardinal and scalar properties, to set robust baselines, and to asses if poverty is changing over time to meet SDG targets. This paper considers the underlying methodologies of approaches that are in place to meet this challenge for multi-dimensional child poverty, both in terms of methods and in terms of primary data sources.

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

Data for the SDGs for multi-dimensional poverty measurement will be a mix of existing surveys that were developed prior to the Millennium Development Goals (MDG) era, and new data, both survey and administrative/'big data'. Many of the specific surveys developed prior to or for MDGs in the developing world will continue and will adapt to incorporate new measures. However, many of these adaptations will be to include new SDG goal areas or targets, and may not adapted to be have elements that can be considered together as a set of indicators to optimally capture multi-dimensional poverty. In this paper 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

Child Multidimensional Poverty Indices

We limit our discussion to 'counting indices' drawn from DHS and MICS survey data. In such indices indicators are arithmetically summed according to a range of different weighting and aggregation assumptions. But, crucially, no statistical derivation of probabilistic weights or relationships are assigned to the indicators in such indices. 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. This 'sum-count' index has no necessarily explicit reference to theoretically or normatively set 'dimensions' that can assign indicators to a typology of poverty need areas. We use such a simple 'sum-count' index in our discussion and analysis below as a counterfactual benchmark index for comparison to the main indices under analysis.

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

- Alkire Foster methodology is an index formed from a sum of differently weighted indicators. Weights are determined according the assignment of dimensions for classifying indicators. The sum of dimensions will be 1, but each indicator within a dimension will have a weight determined by the number of indicators in that dimension. In the most well know precedent, the global MPI, three dimensions were set to reflect the UNDP's Human Development Index (health, education and living standards) so that each dimension had a weight of 1/3. The MPI has two indicators in each of the education and health dimensions, resulting in indicator weights of 1/6; while the 6 indicators in the living standards' dimension, results in indicators weights of 1/18 for those indicators. But, while the MPI dominates discussion, it should be considered a variant of Alkire Foster, 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. Also, a crucially important clarification is that *dimension weights in Alkire Foster need not axiomatically set be set equally*, as in MPI. For instance, this means that dimensional weighting can consider important empirical considerations, such as clustering of household indicators, and important potential consideration for SDG measurement of individual level child poverty that we return to discuss later.

  The headcount measure for an Alkire Foster index is then accompanied by 'intensity' and 'adjusted headcount' measures that allow a complete reconciliation of poverty measurement to the Foster, Greer Thorbecke 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). The algebraic proof for 'Alkire Foster' is given in the full version of this paper.

  The dominance of global MPI as a 'brand' of Alkire Foster has arisen since 2010 as the UNDP Human Development Report was instrumental in its launch and establishing the methodology in global poverty practice. But the global MPI is not 'Alkire Foster'; just one version of the methodology. National level 'MPIs' adopted by governments all over the developing world, differ from the global version in many ways. Children's multi-dimensional poverty can also be measured by disaggregating household level MPI – as most recently done at the global level for the first time (Alkire et al 2017). Child level Alkire Foster measures were also established early in the literature (Roche 2013) but they 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. This term is ours and refers to a number of indices that use a normative ‘rights based approach’ to construct an index that sums *dimensions*, which are populated by groups of indicators[2]. The crucial arithmetic differences to both Alkire Foster 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

  These indices are of longer standing – starting with Gordon et al’s 2003 global child poverty profile (op-cit), then put in place under UNICEF’s global poverty and disparities profiles in around 50 countries, and by ECLAC in Latin America regional and national child poverty proofing (CEPAL/UNCEF 2010, 2012), and, most recently, by the Multiple Overlapping Deprivation Analysis (MODA) (De Neuburgh et al 2013, others). Some countries have also adopted this rights-based normative approach for national multidimensional indices. The algebraic formula and proof for categorical counting approach is given in the full version of this paper.

The Analysis

How is it best to compare these approaches? 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). Clear differences in the level and composition of poverty are often found in these studies, but such differences can be difficult to interpret if it is not clear what arises from the underlying methodology, or from the construction in survey data (through use of different indicators, difference in the construction of similar indicators, and the underlying data cleaning work such as trimming of data for outliers, etc.). There is a much larger analytical literature on the Alkire Foster approach and in its comparison to statistical and econometric measurement practice of poverty in general. For instance, tests of robustness and sensitivity for MPI are undertaken (Alkire 2014, for instance), alternative theoretical measurement approaches compared (Rippin 2010 and others). 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). Technical evaluation of the categorical counting approach has been minimal by comparison.

[2] Another term could be ‘dimensional counting’ but with Alkire Foster decomposition often producing results by dimension, we considered out term less ambiguous.

Our primary research questions go back to the applied question of poverty measurement for the SDGs: How do the Alkire Foster and Categorical Counting 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 analytical approach is a cumulative one to consider the properties that arise by construction, from axiomatic measurement principles in the first instance and then by assessing how these properties affect performance in actual household survey data. To do so, we do not start with an algebraic proof for several reasons. First, our aim is to reach an audience that is wider and 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 different assumptions – for instance, that the Greek symbol sigma Σ, for 'sum', may hide a cumulative sum of non-consistent underlying components (and thus differences between ordinal, categorical and cardinal numbers). Third, we use worked examples from 'laboratory data' and these include the outcomes of 'trial and error' in some instances. It is sometimes difficult to demonstrate theoretical measurement problems foreseen by mathematical proofs in the lab, and adjusting and reconfiguring laboratory data to solve these problems can tell us more about the performance of indices than simple theoretical assertions. However, algebraic proofs are extremely useful and essential reference material and we include them in Appendices in the full version of this paper.

We then move from laboratory data to implement the indices in real survey data and thus to move from discussions of measurement 'theory' to 'practice'. We test real micro-survey data from three countries to see if the findings from laboratory tests are validated.

Our motivation is to concentrate on underlying methodology and its applied outcomes for data and poverty profiles. We want to go beyond simple descriptive comparisons of indices already in place. Throughout our analysis we thus avoid 'brand' comparisons of particular named indices but instead concentrate on the underlying measurement approach and its assumptions. We recognize the investments of many kinds that have gone into different particular named indices but our analysis is of measurement methodology – of Alkire Foster and Categorical Counting approaches, and not of the indices that spring from them:, MODA, MPI, CEPAL/UNICEF etc.

Part 1: Laboratory Data and Multidimensional Counting Indices

Laboratory data is constructed using 10,000 hypothetical observations each allocated 10 non-specified indicators. We randomly (coin-flip) allocate binary ones and zero scores to each indicator. This means that all indicators are independent of each other and there is, by definition, no correlation between them, a factor that we reflect on further in our sensitivity analysis. To make comparisons between indices we use simulation from 100 Monte Carlo trials of coin-flip random allocation, and with those results that we have distributions that produce 50% mean multiple deprivation to ensure correct random assignment of indicator status (1, 0) to have a known arithmetic outcome of mean multiple deprivation of 50%.

We construct the multi-dimensional indices using this dataset of randomly assigned 10 deprivations for 10,000 observations. We construct Alkire Foster index using the same weights as those discussed above – reflecting the most well-known version of their methodology – the global MPI. Four of the ten indicators have weights of 1/6 and six indicators have weights of 1/18. This replicates for children the form of individual level MPI index demonstrated by Klasen and Lahoti (2016). We construct the Categorical Counting index by also using approaches from the applied literature (Gordon et al 2003 and De Neuburg et al 2013) to have five dimensions: three dimensions are populated by 2 indicators in 'union approach', one dimension has 3 indicators in union approach and the fifth dimension has just one indicator. For comparison purposes, we also calculate a 'sum-count' index of the same distribution for comparison – as a counterfactual where no weighting or aggregation principles are employed (beyond simple equal weights for each indicator). All versions are subject to Monte Carlo 100 simulation trials to provide robust estimates of results at 99.9 percent level.

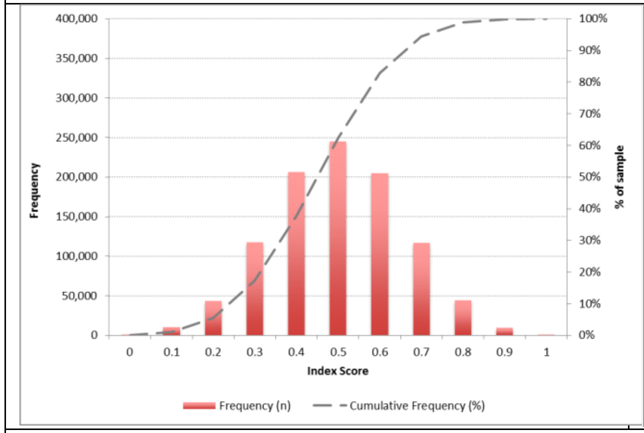
Figure 1 shows the results. Figure 1 a) shows the 'sum count' benchmark index, which, definition and by construction, produces a mean and median score of 0.5. It has a normal distribution that has 11 scalar increments set at 0.1 apart between zero and one, and that has and a standard deviation of 0.16.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

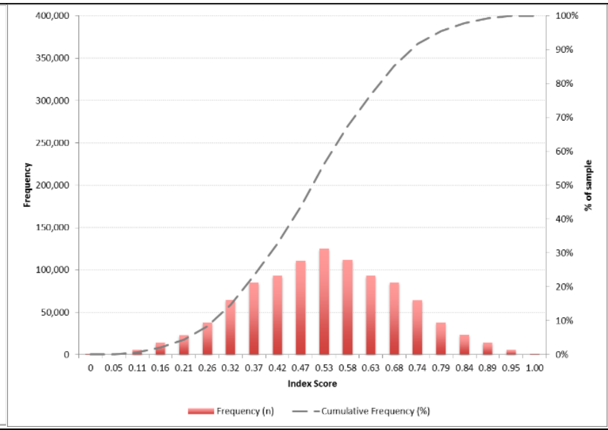
Figure 1b) shows the Alkire Foster MPI specification also results visually in a bell curve distribution, (but Shapiro-Wilk tests do *not confirm* it is a normal distribution) with 18 equal increments of 0.566 (the value of the smallest weight) between 0 and 1. Having index weights smaller than 1/10 results 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.

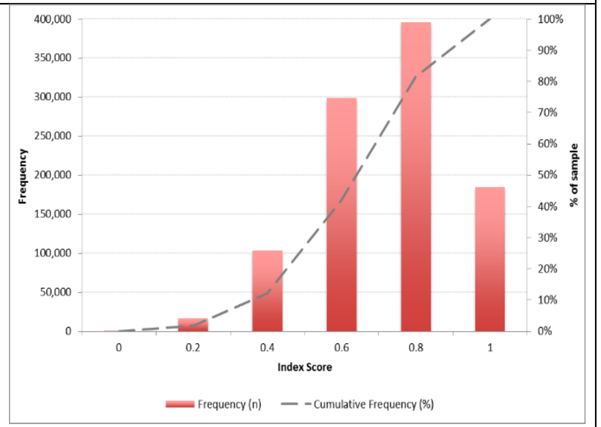
Figure 1.c) shows the results for the Categorical Counting index, which stands in stark contrast to both the benchmark and Alkire Foster specifications for the same data and underlying distribution of deprivation scores. The first thing to note is the far less granular distribution, as counting 'dimensions' 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 Categorical Counting approach is hugely skewed, with a skewed tail to the left, (and resulting negative Skewness score of 0.4 compared to scores of close to zero for the other specifications) and thus a skewed peak of the distribution to the right, and thus of resulting higher index scores overall: a mean of 0.73 and a median of 0.8. It is worth repeating and reassuring readers that these results come from the same population distribution that itself came from the same random assignment of 10 deprivation scores by coin-flip that derived the other 'sum count' and Alkire Foster results. Simply said, the Categorical Counting specification 'exaggerates' 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 multi-dimensional methodologies using the 'union approach' result in 'exaggeration' of poverty estimates. 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 Categorical Counting verses the other indices is a property that must be explicitly addressed when assessing such baselines in practice.

Figure 1 Baseline Results
Monte Carlo Simulations (100 trials)

a) 'Sum-Count'- Benchmark	b) Alkire Foster (Global MPI weights)	c) Categorical Counting (5 dimensions, 1x1, 3x2, 1x3)







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

Of course, we are mindful that particular specifications of Alkire Foster or Categorical Counting 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 Categorical Counting exaggerates poverty at all thresholds compared to the Alkire Foster and 'sum count' specification in any form. Confirmatory results can be obtained from the authors.

Household Clustering, Age-Specific Censoring and Saturation

We use the laboratory dataset to consider the differences that will occur as a result of the two measurement problems we identified earlier when discussing survey data and individual child level indices,

- some indicators are clustered at the household level and others are at the individual level
- individual level indicators are age specific and any individual outside of that age-banding is censored for that indicator

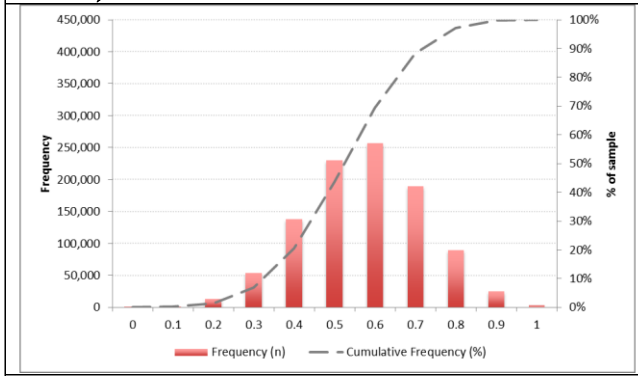
How do the distributions described in Figure 1 change if some of the observations and indicators are at the household level? We reconfigured the laboratory dataset to assign individual observations to 'pseudo households' randomly. We set the average number of observations per pseudo household as 3, while allowing up to a maximum of 7 observations to be allocated to a random household. Pseudo household variable's summary statistics is presented in Appendix B. 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'. The exercise is designed to illustrate the potential effects of household clustering compared to a random allocation and is not intended to be representative of actual household formation and size.

These results are clearly seen in Figure 2 with all three distributions more skewed than their original individual level baselines in Figure 1. The sensitivity of indices to the randomized allocation of observations to households demonstrates that the Categorical Counting index reaches its ceiling value relatively quickly, at which point is 'saturated' and less irresponsive to any incremental positive changes in any of its indicators. On the other hand, as the Sum Count and Alkire Foster 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 (all observations who share the same pseudo household now have the same values for the two household level indicators). There is an increase in overall deprivation scores in households with more household members as the ratio of ones to zeros changes to favor the former.

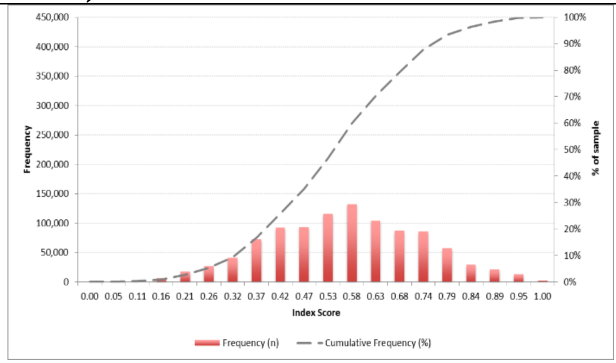
Compared to Figure 1, skewness scores turn negative for the sum-count and Alkire Foster specifications and increase in skewness for Categorical Counting. By construction, standard deviation scores fall as the result of having more repeated values in the overall distribution. The mean index scores increase by statistically significant levels (at 1% level) across all specifications. However, the increases are different in size: the mean rises from 0.50 to 0.57 for the benchmark, but from 0.50 to 0.54 for Alkire Foster, a reflection that on average *lower weighted indicators were affected*. This raises the possibility that Alkire Foster type weights can be used to address household clustering effects on individual level indices. We return to this point later in discussion. On the other hand, the Categorical Counting index increased its mean score from 0.73 to 0.82 – a large 0.9 rise in score that reflects greater a resulting weight

Figure 2 Revised results allowing for Household Clustering (2/10 indicators at household level)
Monte Carlo Simulations (100 trials)

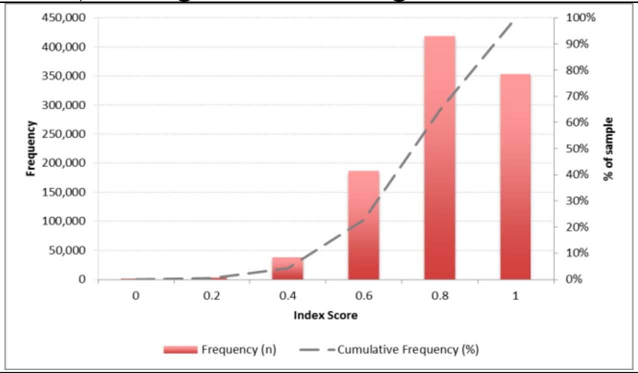
a) Sum-Count Benchmark



b) Alkire Foster[3]



c) Categorical Counting[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	

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)

.
being given to the new household level variables from the union aggregation and dimension level counting. The differential impact of the same household clustering on the Alkire Foster and Categorical Counting indices has thus potentially important ramifications for applied use of the indices and in their interpretation.

Introducing two household level variables into our laboratory dataset allows us to consider the impact of household clustering on 2 out of 10 indicators. However, when we consider the practice of Categorical Counting child poverty indices (Gordon et al 2003, de Neuburgh et al 2013), we see that household level indicators are a much higher proportion of all indicators – often eight out of ten indicators will be at the household level.

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

	Sum-Count Benchmark	**Alkire Foster[5]**	**Categorical Counting [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

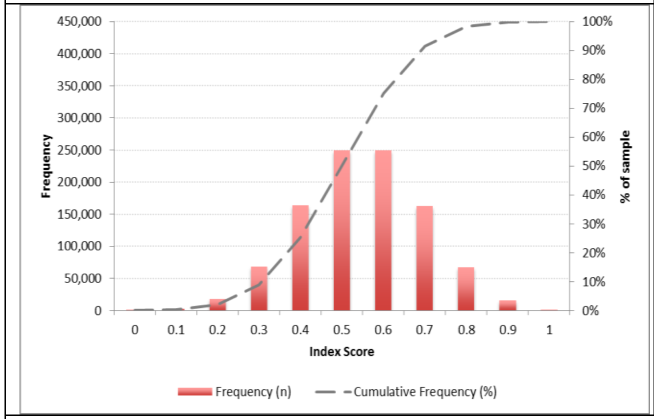
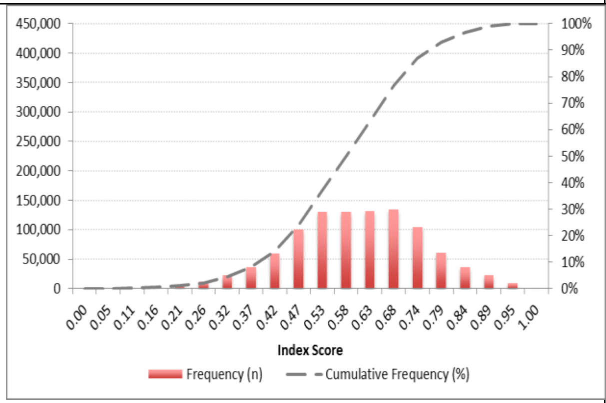
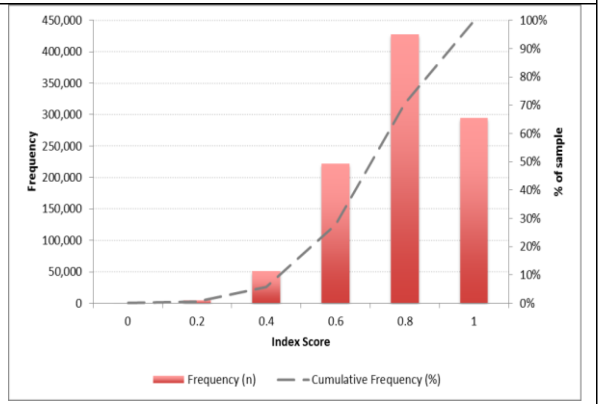
Table 1 gives an indication of the difference to Figure 2's result's that would from such higher levels of household clustering. The Categorical Counting index's saturation is now clear, with mean at 0.93 and median at 1, a caveat for indices of this type in contexts with higher levels of deprivation and high proportion of indicators at household level. On the other hand, Alkire Foster 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 additionally, and cumulatively, at age specific censoring and its potential impact on our laboratory data. By definition only individual level variables can be age-specific. If individual level indicators are thus censored (because they are derived as missing -not observed for this age-group), the ratio of observed individual level variance falls relative to household level repeated values and the potential for the dominance of the index from clustered household level variables rises. To capture this effect we do a further transformation of the household clustered version of our laboratory dataset that was seen in Figure 2. We censored 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 countries (UNDESA 2015) that we consider later in Part 2. To do this we changed all 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 re-weight to reflect underlying population differences. However, in applied use of indices, population reweighting would require more consideration. We return to discuss this issue of weighting in Part 2.

Figure 3 gives the revised results for this transformation and its effect on the indices. We show the net change in summary statistics that are observed compared to Figure 5. However,

[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)

Figure 3 Revised results allowing for Household Clustering and Age-Specific Incidence of Deprivation
Monte Carlo Simulations (100 trials)

Multiple Deprivation- Benchmark		**Alkire Foster**		**Categorical Counting**	
					
Summary Statistical Properties & Differences from Iteration shown in Figure 5					
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	

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

we know that censoring in this form is exactly the same as our later tests of sensitivity through changing prevalence in indicators. Out later discussion of sensitivity can thus help interpret some of the results, but we return to discuss this in the next section.

Figure 3 shows the 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 2 results. For Alkire Foster 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. On the other hand the Categorical Counting specification shows a small significant (at 1%) decrease in mean score and an increased in standard deviation. However, the underlying Skewness measure has risen to the highest level across all three simulations across Figure 1, 2 and 3. Taken together these results show real inconsistencies in the way that the Categorical Counting index reacts differently to changes in individual level prevalence in variables compared to the Sum Count and Alkire Foster indices and these, in part, can be interpreted as deriving from the different properties of monotonicity explored in the next section of the paper. We are thus able to make initial findings with a high degree of certainty in the laboratory about the first of our comparison questions on the differences in cardinal and scalar properties of the indices.

Sensitivity Tests 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 they asses if poverty is changing over time to meet SDG targets? This is a key question for poverty measurement for both accurately identifying difference between sub-groups and for tracking change over time. The use of 'flip coin' random assignment to create indicator prevalence (and validating Monte Carlo trials) produces indicator relationships that are, by definition, random and not correlated. This introduces a limitation to the scope and interpretation of results that we can generate using this approach and we will also consider how correlation between indicators affects sensitivity and monotonicity in our analysis.

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 Alkire Foster 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 Categorical Counting 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 Categorical Counting index is asymmetric.

Figure 4
Changes in Index Scores from Incremental Change to Any Indicator
(100 Monte Carlo Trials)

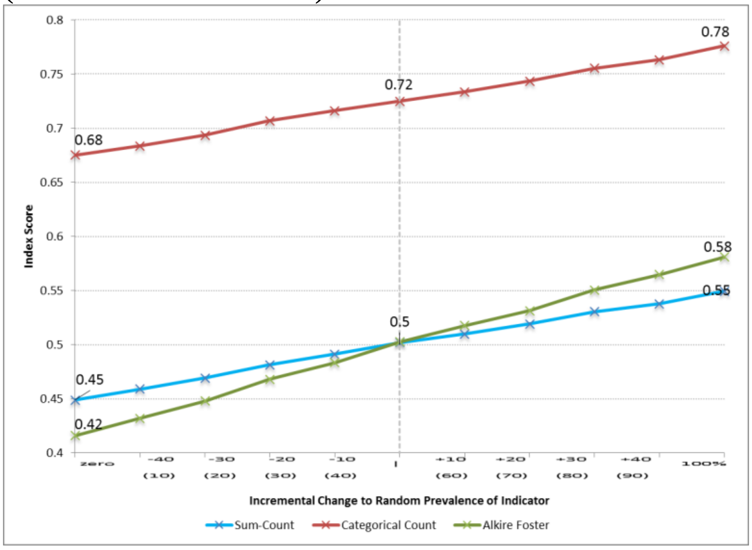
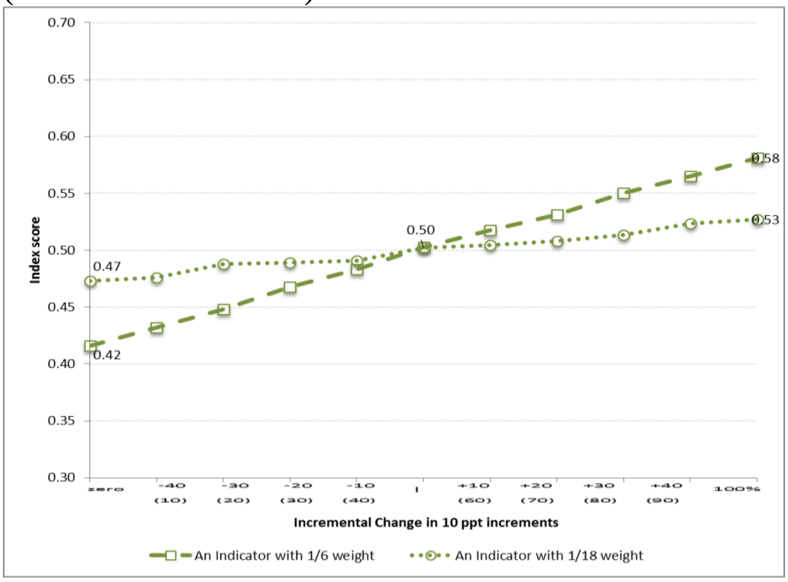


Figure 5 shows the influence of differential weights in the Alkire Foster index. The level of incremental change differs: large change is the result of higher indicator weight, but changes similarly symmetric across incremental change to indicators of differential weight.

Figure 5
Alkire Foster: Incremental Change in Indicator by Assigned Weight
(100 Monte Carlo trials)



When we attempt to replicated Figure 5 for the Categorical Counting 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 has two effects for the Categorical Counting index:

- first, saturation, that will prevent the index from representing 'increased' poverty risk from increased prevalence in indicators in the same way as its 'unsaturated' form; and

- second, the actual drivers of sensitivity for the Categorical Index will not be limited to changes in prevalence of an indicator but also to its correlation with indicators that are used in ‘union’ to populate the categorical dimension for counting

We tackle these two issues in turn. Figure 6 shows the 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

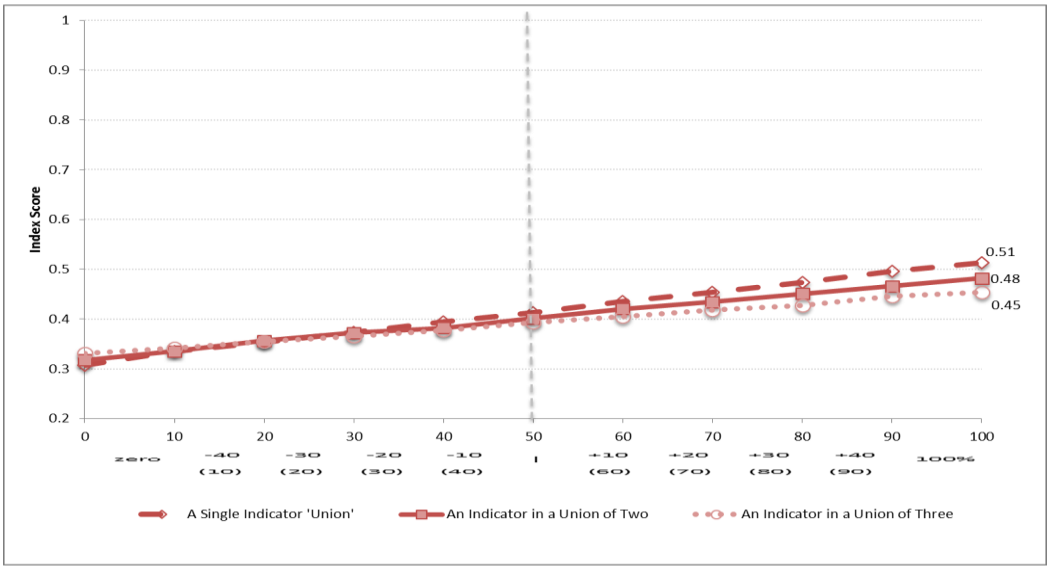
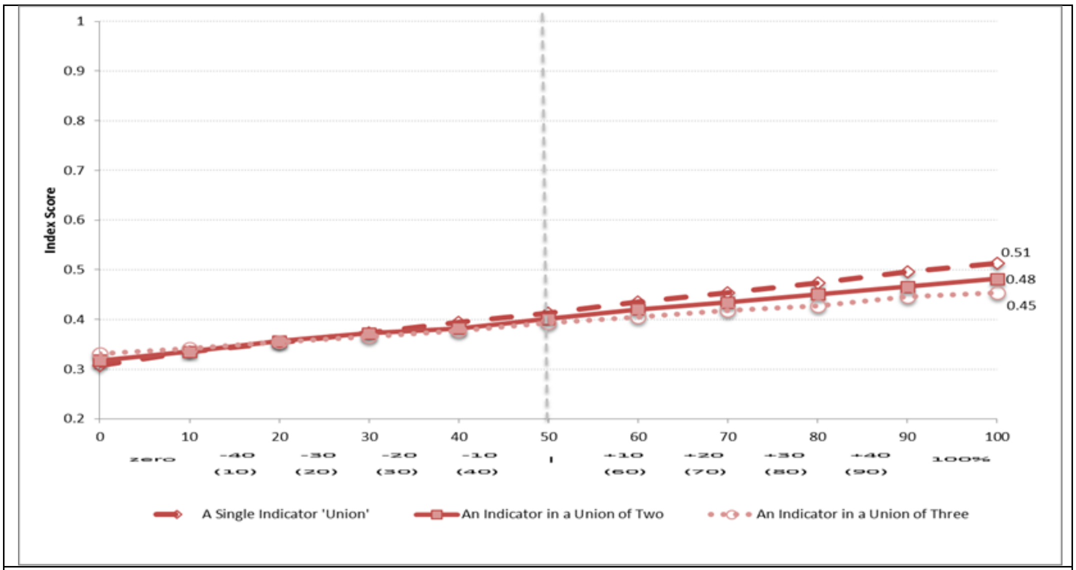
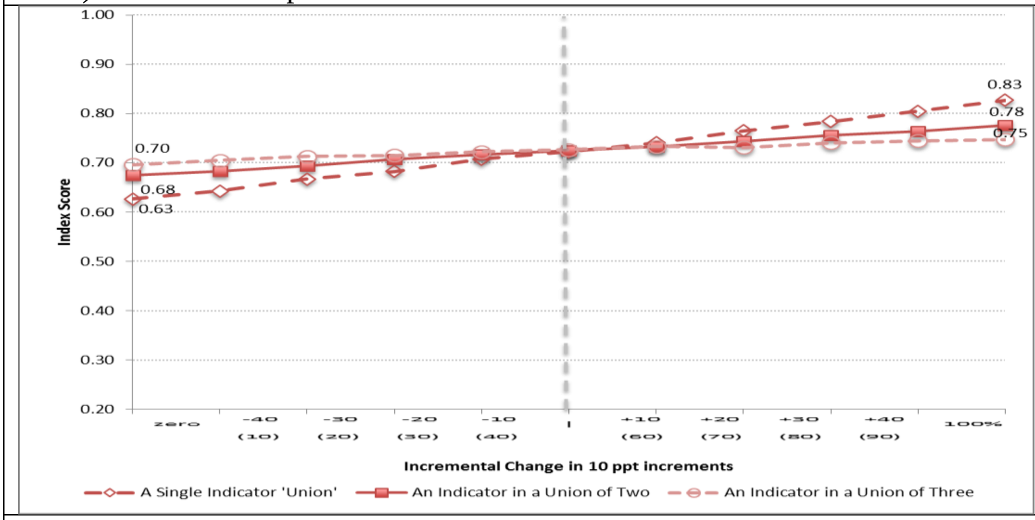


Figure 6
Categorical Counting: Incremental Change in Indicator by Union Assumption with 20%, 50% and 80% random prevalence
(100 Monte Carlo Trials)

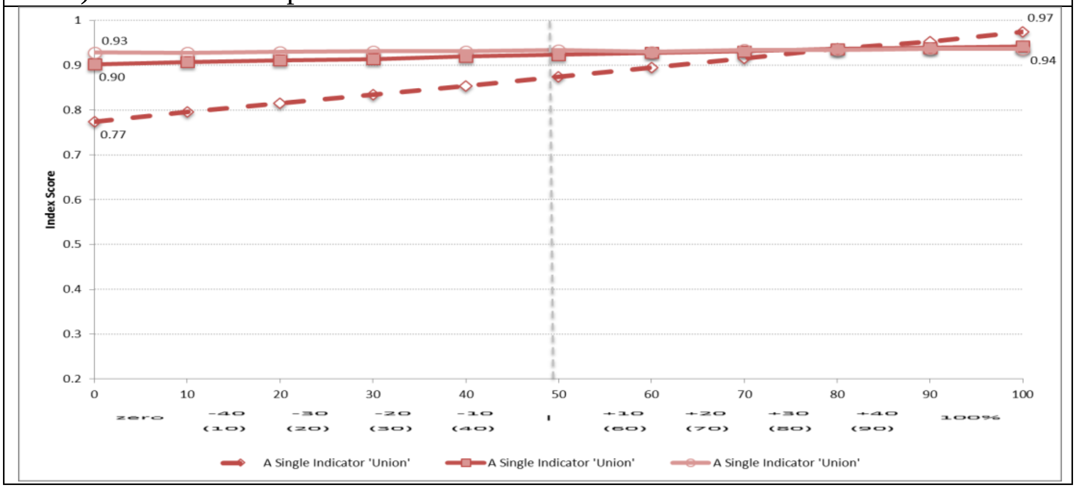
a) 20% random prevalence



b) 50% random prevalence



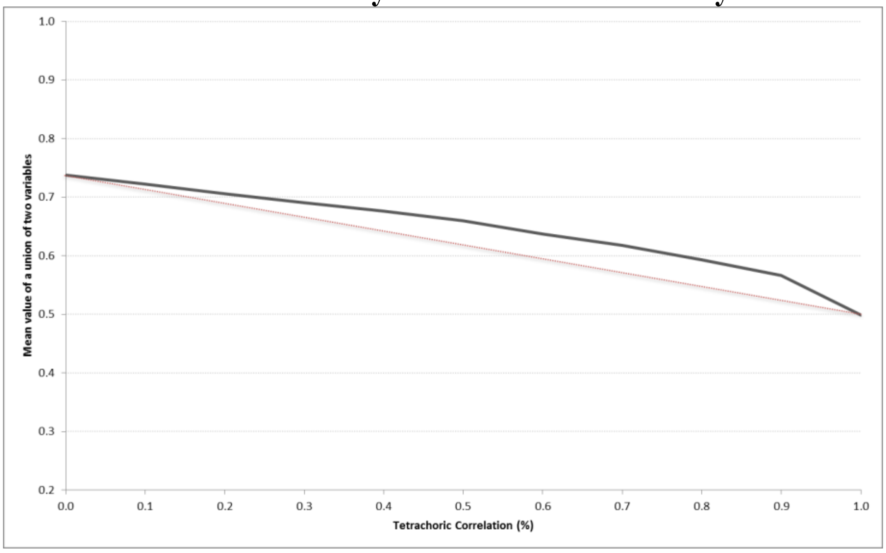
c) 80% random prevalence



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'.

But random allocation of indicator prevalence at any level still results in indicators not being correlated, which for the Categorical Counting index, is a major limitation to testing sensitivity and monotonicity, due to the inherent differences in probability by its union assumptions for construction and its resulting sensitivity to correlation between indicators aggregated in union. First, we demonstrate the underlying difference that occurs by virtue of correlation in the mean value of indicators placed in union. Figure 7 shows the changing mean of two indicators in union by the level of correlation between them

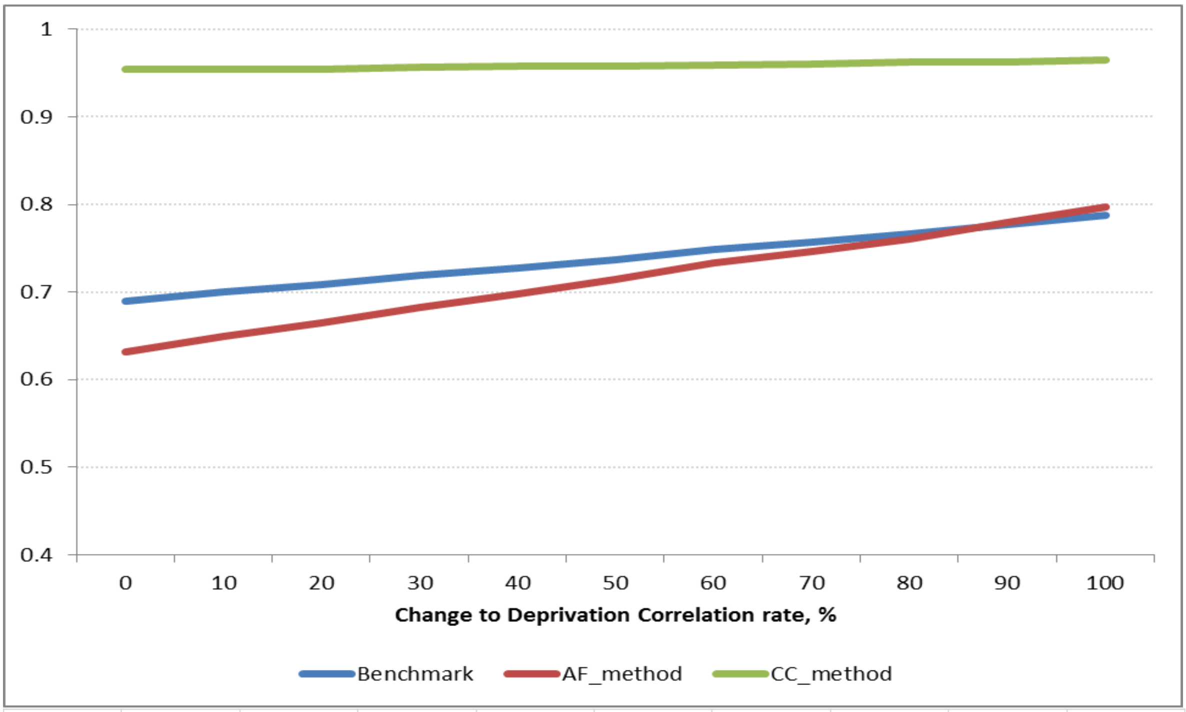
Figure 7
Mean Values of Two Binary Indicators in "Union' by Level of Correlation



We see that high levels of correlation would give non-linear profile to mean values for each 'counted' dimension produced using the 'union assumption' in the Categorical Count index.

Figure 8 shows how a hypothetical set of correlations between 10 indicators alters the performance of all three indices. Based on our findings from Figure 7 and the higher likelihood of asymmetry for more highly correlated indicators, we demonstrate how the indices would follow a dominant indicator and the potential for information from the other indicators to be 'lost' due to their high correlation with the dominant indicator. Figure 8 shows that the Alkire Foster index converges with the Sum Count at high levels of correlation but that both these indices still show increases in score as correlation increases. On the other hand the Categorical Counting index loses ability to capture change due to higher correlation, and it thus far more sensitive to correlation that the other indices.

Figure 8
Sensitivity of Counting Indices to High Correlation to Dominant Indicator



	Indicator A	Indicator B	Indicator C	Indicator D	Indicator E	Indicator F	Indicator G	Indicator H	Indicator I
Indicator A	A								
Indicator B	0.0112	B							
Indicator C	0.0121	0.2138	C						
Indicator D	0.0051	-0.019	0.0021	D					
Indicator E	-0.0001	0.0043	0.3117	-0.0045	E				
Indicator F	0.0013	-0.0049	-0.0106	-0.0039	0.2787	F			
Indicator G	0.0021	-0.0031	-0.0147	-0.0096	0.0076	-0.0105	G		
Indicator H	0.0025	0.0108	0.4688	-0.006	0.2535	-0.0053	-0.0108	H	
Indicator I	0.0144	0.0222	0.7773	-0.0037	0.4139	-0.0093	-0.0073	0.6103	I

Overall our sensitivity analysis suggests that Alkire Foster has monotonic properties that differ by the assigned weight given to any indicator. Categorical Counting has problematic monotonicity due to non-consistently different impacts of indicator change, that can vary in size due to implicit rather than explicit weights, and that can be asymmetric depending on how the indicator that changes probability of effecting the index score arises from its assignment in union or not to a dimension. This gives us a set of findings on our third question on the 'monotonicity' of the indices. Algebraic proofs to demonstrate these characteristics and support our laboratory findings are given in the full version of this paper.

Part 2: Indices using Household Survey Data

In the second part of our analysis we take forward the key lessons from the laboratory work and test them with real survey data. At this point it is crucial that we restate our motivation as our work will NOT replicate actual indices that are in place in these countries or used in any form. We repeat our earlier clarification of the aim of this paper: to test underlying methodologies not to test or compare 'branded indices' 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 similar 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 Cambodia, Bangladesh and Tanzania. We take 10 indicators and construct three indices from those indicators:

- the 'sum-count' benchmark index that has been so useful in the lab work to interpret differences between it and the other indices.
- Alkire Foster: we have 3 dimensions (2,3 and 5 indicators per dimension)
- Categorical Counting: 5 dimensions (2,3,1,1,3 indicators per dimension)

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. For this conference version of the paper, we do not consider the clustering and age-specific censoring issues inherent in correlation and reporting results. Readers are pointed to the full version of the paper for these.

Table 2
Indicators and Dimensions for Indices

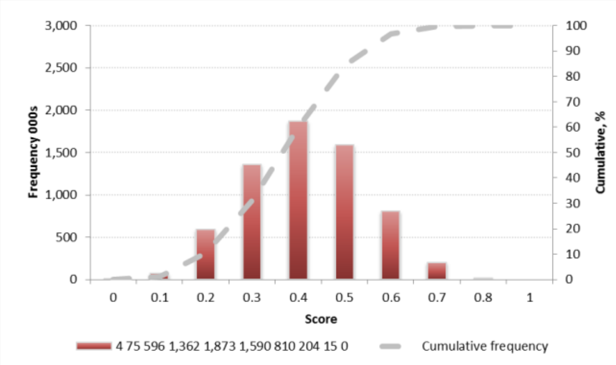
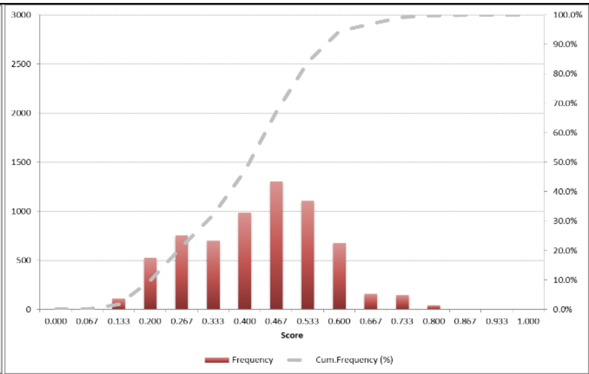
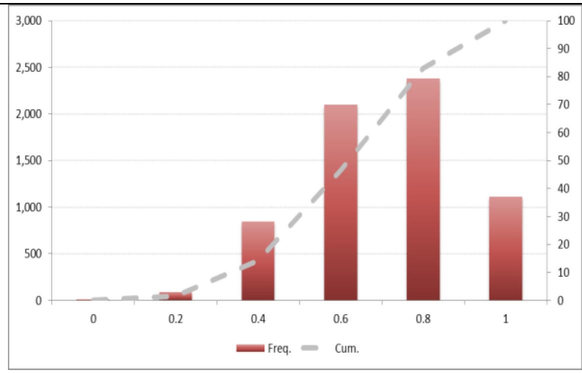
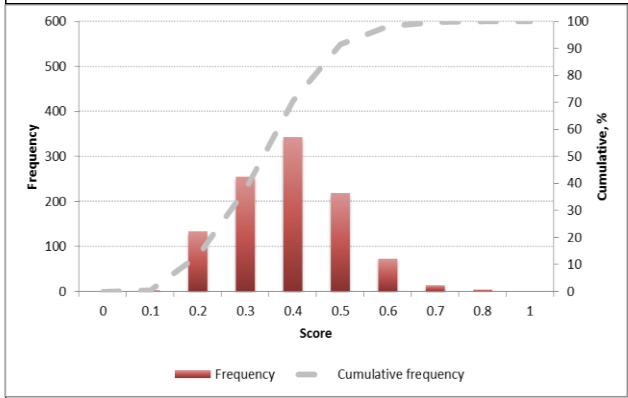
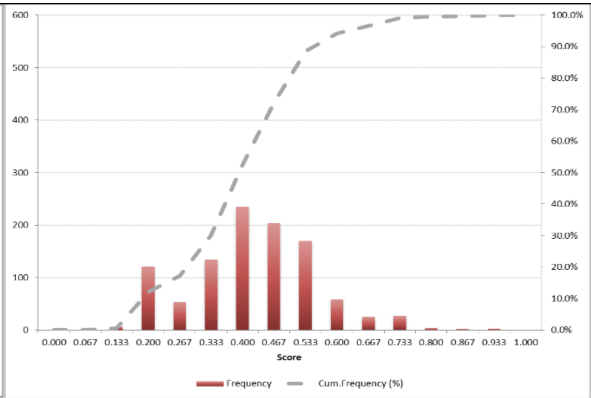
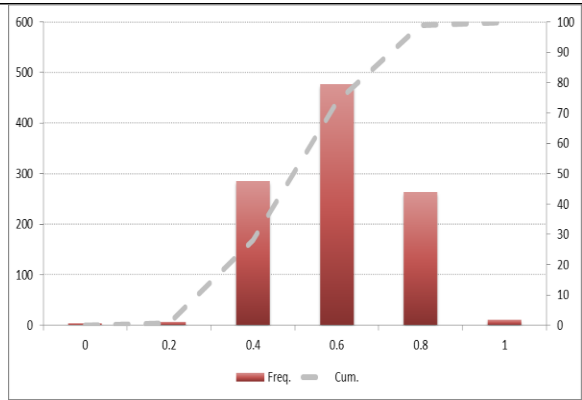
Categorical Counting: **Composition of Dimensions**					
Nutrition	Infant feeding	Wasting			
Health	DPT all	Unskilled birth attendant	Child mortality		
Water	Drinking water				
Sanitation	Toilet type				
LS	Overcrowding	Wealth low quintile	Info devices		
AF method: **Composition of Dimensions**					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

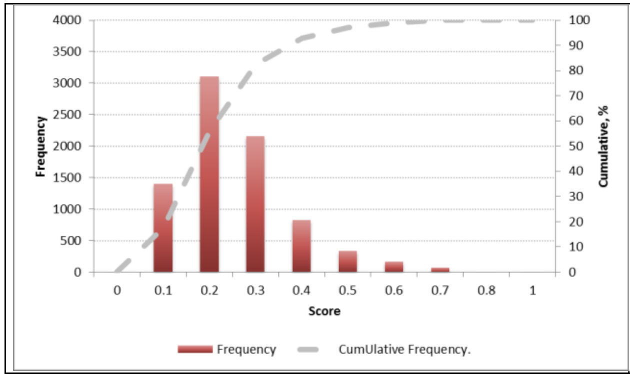
We do not test our set of indicators for their suitability, reliability, validity or underlying robustness in their performance for any overall index specification. We are not interested in how these indices accurately assess multi-dimensional poverty, merely in their comparative performance for assessment of methodology according to underlying measurement properties. 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, but to capture a consistent set of deprivations to construct indices for comparison of measurement properties, as a follow on from our earlier lab testing. 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. However, we do assess the correlations between indicators to enable us to consider correlation as an important factor in performance of the indices, as outlined in the previous section. Table 3 shows the correlation matrices for the ten named indicators for each data set.

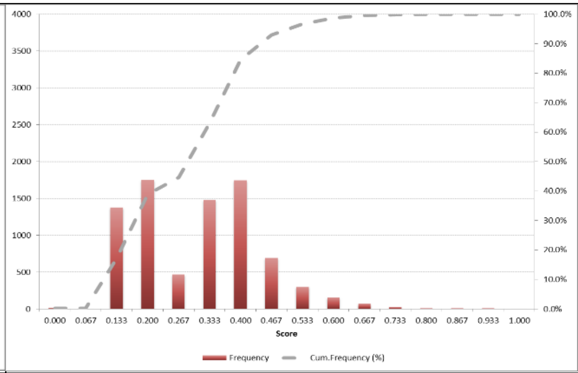
Table 3
Tetrachoric Correlation of 10 Chosen Indicators

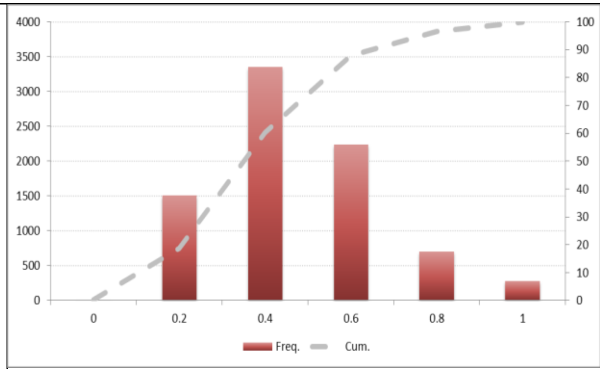
		Drinking water	Toilet type	Overcrowding	Infant feeding	Wasting	DPT all	Unskilled birth attendant	HH level Child mortality	Wealth low quintile	Info devices
CAMBODIA	Drinking water	1									
	Toilet type	0.66	1								
	Overcrowding	0.33	0.37	1							
	Infant feeding	0.16	0.23	0.11	1						
	Wasting	0.01	0.07	0.17	-0.08	1					
	DPT all	0.12	0.26	0.12	0.19	0.02	1				
	Unskilled birth attendant	0.54	0.54	0.36	0.20	0.11	0.31	1			
	HH level Child mortality	-0.16	-0.22	-0.09	-0.11	-0.05	-0.09	-0.14	1		
	Wealth low quintile	-0.47	-0.65	-0.43	-0.17	-0.03	-0.16	-0.49	0.17	1	
	Info devices	0.55	0.60	0.32	0.19	0.11	0.19	0.60	-0.06	-1.00	1
BANGLADESH	Drinking water	1									
	Toilet type	-0.01	1								
	Overcrowding	0.22	0.27	1							
	Infant feeding	-0.07	0.06	0.10	1						
	Wasting	0.11	0.15	-0.01	-0.14	1					
	DPT all	0.40	0.02	0.13	0.12	0.10	1				
	Unskilled birth attendant	0.10	0.42	0.35	0.13	0.07	0.27	1			
	HH level Child mortality	-0.02	-0.18	0.06	-0.08	0.21	-0.13	-0.26	1		
	Wealth low quintile	-0.14	-0.52	-0.34	-0.10	-0.05	-0.16	-0.56	0.20	1	
	Info devices	0.08	0.47	0.42	0.07	0.03	0.18	0.53	-0.04	-0.98	1
TANZANIA	Drinking water	1									
	Toilet type	0.44	1								
	Overcrowding	0.06	0.14	1							
	Infant feeding	0.12	0.05	0.03	1						
	Wasting	-0.03	-0.11	0.01	0.09	1					
	DPT all	0.24	0.21	0.13	0.18	-0.08	1				
	Unskilled birth att	0.27	0.45	0.19	0.10	0.08	0.27	1			
	HH level Child mo	-0.07	-0.07	0.05	-0.06	0.00	-0.13	-0.16	1		
	Wealth low quinti	-0.31	-0.48	-0.13	-0.03	0.00	-0.12	-0.32	-0.01	1	
	Info devices	0.24	0.36	0.13	-0.01	-0.01	0.18	0.23	0.04	-0.95	1

Figure 7
Headline Results for Indices in Three Counties: distributions, means and maximum values

<table>
<tr><th colspan="2">a) ‘Sum-Count’</th><th colspan="2">b) Alkire Foster</th><th colspan="2">c) Categorical Counting</th></tr>
<tr><td colspan="6">TANZANIA</td></tr>
<tr><td colspan="2">
</td>
<td colspan="2">
</td>
<td colspan="2">
</td></tr>
<tr><td>Mean</td><td>0.42</td><td>Mean</td><td>0.40</td><td>Mean</td><td>0.72</td></tr>
<tr><td>Maximum</td><td>0.8.0</td><td>Maximum</td><td>0.87</td><td>Median</td><td>1.00</td></tr>
<tr><td colspan="6">BANGLADESH</td></tr>
<tr><td colspan="2">
</td>
<td colspan="2">
</td>
<td colspan="2">
</td></tr>
<tr><td>Mean</td><td>0.40</td><td>Mean</td><td>0.40</td><td>Mean</td><td>0.61</td></tr>
<tr><td>Maximum</td><td>0.80</td><td>Maximum</td><td>0.87</td><td>Maximum</td><td>1.00</td></tr>
<tr><td colspan="6">CAMBODIA</td></tr>
</table>







Mean	0.23	*Mean*	0.25	*Mean*	0.44
Maximum	0.80	*Maximum*	0.87	*Maximum*	1.0

Our three key questions remain uppermost in mind but are now tested to confirm or revise our findings from laboratory work.

Figure 7 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 Categorical Counting and the other indices using actual survey data. The Sum Count and Alkire Foster 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 Alkire Foster. However, the Categorical Counting index gives consistently higher mean scores compared to the other indices – in the region of 50 per cent higher. Additionally, the maximum score for Categorical Counting 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 on robustness of baselines. We find continued results that suggest 'exaggeration' for Categorical Counting indices verses Alkire Foster and the benchmark 'sum count'.

How do the indices perform when considering changes to indicator prevalence? We show illustrative sensitivity tests for two indicators: water and the presence of skilled birth attendant. We have chosen these two to present 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 Categorical Counting 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 Categorical Counting 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.

For this paper, we limit analysis to these two tests to reflect our laboratory work. A fuller set of sensitivity tests can be obtained from the authors. The results from these two sensitivity tests shown in Figure 8. We see common findings in the changes to both indicators from incremental changes in prevalence across countries and across indices. Incremental changes to prevalence of the water indicator has big impact on the Categorical Counting 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 Alkire Foster 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 Categorical Counting index – as shown by the 'flat line' in Figure 8, 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 sensitivity 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 Categorical Counting method. These findings confirm what we saw in the laboratory analysis on results for tests of monotonicity.

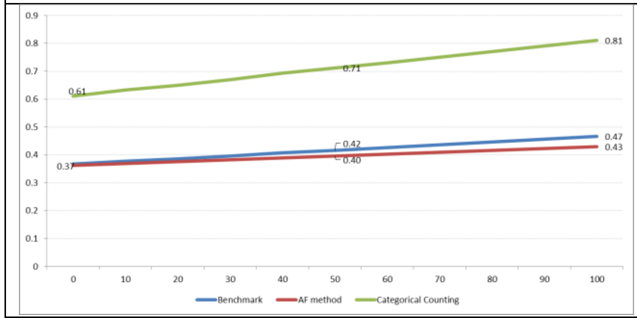
.

Figure 8
Sensitivity Tests for Water and Presence of Skilled birth Attendant Indicators

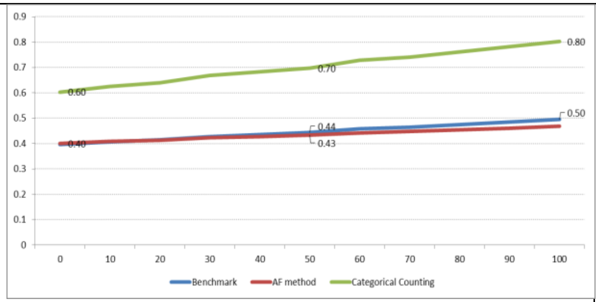
TANZANIA	BANGLADESH	CAMBODIA

Water

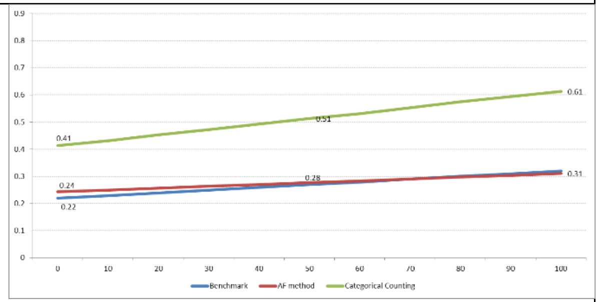
TANZANIA



BANGLADESH

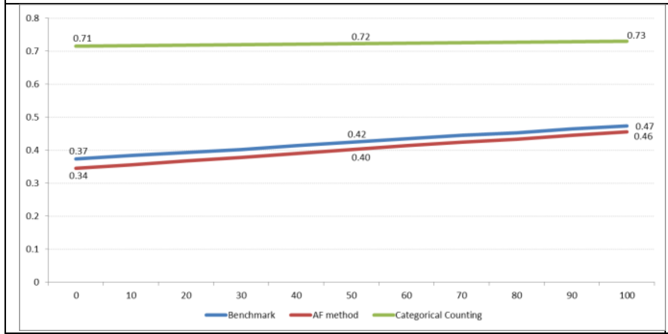


CAMBODIA

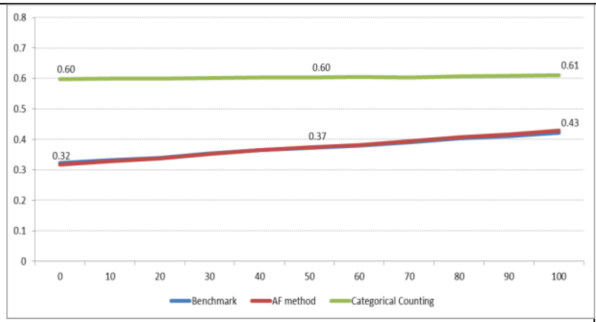


Presence of Skilled Attendant at Birth

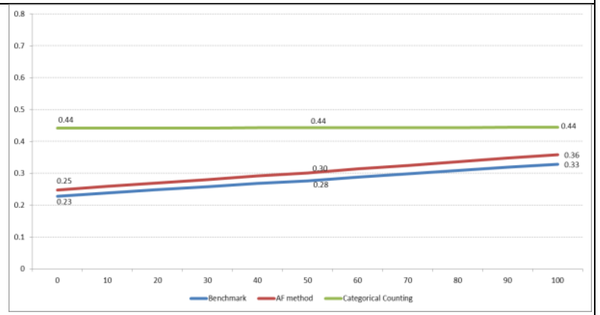
TANZANIA



BANGLADESH



CAMBODIA



However, with real survey data we are also able to see the performance of the indices on how they identify and quantify differences in ‘sub-groups, of the population. Such comparisons also allow our focus to be more directly on key applied aspects of poverty measurement by considering how ranking of ‘within country’ sub-groups differs according to the three indices. We restrict our illustrative example to regional differences in poverty, and regional rankings by poverty in Bangladesh and Tanzania. A fuller set of results is available from the authors. Table 5 shows the regional rankings (and baseline poverty scores using the ‘sum count’ approach) for the 6 regions in Bangladesh. All three indices rank Dhaka and Sylhet first and second respectively, and rank Khulna least poor. But the middle of the rankings are not consistent across any index. These results suggest that key issues of resource allocation to reduce or prevent poverty would have to be robustly assessed to support differences in priorities that could arise from different methodologies used for multi-dimensional poverty. Table 5a shows that the six regions of Bangladesh rank 1 and 2 equally across all three indices in the levels of multi-dimensional poverty for the under-fives. However, the ranking of the next 3 regions differs according to the index used, but all three indices agree the 6th rank. Table 5b shows the same result for Tanzania but has 26 sub-national regions/provinces. We identify those index rankings that differ by 3 places or more from the Sum-Count index in grey. Compared to the Sum-Count, Alkire Foster produces three regional ranking difference of 3 or more places, while Categorical Counting produces higher levels of ranking difference, 8 or around a third of all regions.

These results on regional rankings are indicative rather than conclusive but suggest that there are real uncertainties about ‘sub-group’ differences in poverty that result from adopting different indices. The use of MD poverty indices to regionally allocate funds based on multi-dimensional poverty levels would potentially face huge uncertainty, especially when compared to the simple ‘sum count’ approach.

Table 5
Regional Ranking for Multi-dimensional Poverty Index Scores
Multi-dimensionally Poor Children aged under 5

a) **Bangladesh**

Region	Sum Count	Rank Sum Count	AF Rank	CC Rank
Dhaka	0.422	1	1	1
Sylhet	0.401	2	2	2
Barisal	0.393	3	5	4
Chittagong	0.392	4	3	3
Rajshahi	0.389	5	4	5
Khulna	0.363	6	6	6

Difference in 1 Ranking Place shaded grey

b) Tanzania

	Sum Count	Rank Sum 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

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 Alkire Foster 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. Categorical Counting 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. Alkire Foster 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).

Alkire Foster index scores can always be decomposed back to indicator prevalence but Categorical Counting 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 Categorical Counting has established non-consistent aggregation between dimensions, making arithmetic attribution at the indicator level very problematic.

Our laboratory testing also showed that Categorical Counting 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 Categorical Counting 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 Alkire Foster, but we also saw surprisingly different 'implicit' dimension weights in what were normatively assigned 'equal weights' in Categorical Counting index, reflecting the combination of household level indicators and 'union' properties. But differential weighting in Alkire Foster 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 Categorical Counting 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 Categorical Counting – 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 Categorical Counting – 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 Alkire Foster, 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 decomposition of indices rather than in their formulation. Our findings suggest that the measurement of poverty through categorical counting 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 <short>

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

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

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

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.

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

Chakravarty and D'Ambrosio (2006) The measurement of social exclusion, *Review of Income and Wealth* 52(3), 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 http://hdr.undp.org/en/content/multidimensional-poverty-index-achievements-conceptual-and-empirical-issues

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

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

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

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

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.

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, Issue 2, pp 235–248

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.

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

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