Nájera, H. E. (2018). Reliability, Population Classification and Weighting in Multidimensional Poverty Measurement: A Monte Carlo Study. *Social Indicators Research*. https://doi.org/10.1007/s11205-018-1950-z

Social Indicators Research
https://doi.org/10.1007/s11205-018-1950-z

Reliability, Population Classification and Weighting in Multidimensional Poverty Measurement: A Monte Carlo Study

Héctor E. Nájera Catalán[1]

Accepted: 18 June 2018

Abstract
In poverty measurement, differential weighting aims to take into account the unequal importance of the diverse dimensions and aspects of poverty and to add valuable information that improves the classification of the poor and the not-poor. This practice, however, is in contention with both classical test theory and modern measurement theories, which state that high reliability is a necessary condition for consistent population classification, while differential weighting is not so. The literature needs a clear numerical illustration of the relationship between high/low reliability and good/poor population classification to dissolve this tension and assist applied researchers in the assessment of multidimensional poverty indexes, using different reliability statistics. This paper uses a Monte Carlo study based on factor mixture models to draw up a series of uni-and multidimensional poverty measures with different reliabilities and predefined groups. The article shows that low reliability results in a high proportion of the poor group erroneously classified as part of the not poor group. Therefore, reliability inspections should be a systematic practice in poverty measurement. The article provides guidelines for interpreting the effects of unreliability upon adequate population classification and suggest that the classification error of current unreliable multidimensional indexes is above 10%.

Keywords Poverty · Deprivation · Weighting · Reliability · Relative entropy

1 Introduction

Poverty can be defined as the lack of command of resources over time, where different kinds of deprivation are its consequences (Townsend 1979; Gordon 2006). Townsend's (1979) theory predicts that there is a point on the distribution of resources (the Townsend breaking point) from which material and social deprivation (severity of deprivation)

ESRC Project: ES/P009778/1.

✉ Héctor E. Nájera Catalán
pthen@bristol.ac.uk

[1] School for Policy Studies, Centre for the Study of Poverty and Social Justice, University of Bristol, Bristol, United Kingdom

increase substantially. Therefore, this split results in two distinguishable and conceptually meaningful groups—the poor and the not-poor (Gordon 2006). Deprivation mirrors different clusters of unsatisfied human needs and this opens up theoretical and empirical questions about the additive nature of observed deprivations (Gordon 2006; Kakwani and Silber 2008; Alkire and Foster 2011; Guio et al. 2016). Theories on human needs suggest that material and social necessities play diverse roles in people's lives and impact differently on living standards (Sen 1999; Nussbaum 2001; Dean 2010). This raises the question about the best aggregation procedure- inasmuch as the necessities of life contribute differently to an individual's standards of living, it is unclear how these differences should be accounted for, i.e. using equal or unequal weights (Sen 1979; Atkinson 1987).

Any kind of weighting (i.e. equal or differential) attaches a value judgement of a human need's importance relative to others. Assigning different weights to deprivation indicators and dimensions of poverty has been utilised as a means of stressing that some needs are more vital than others (Decancq and Lugo 2013). When looking at the value that populations attach to different social and material needs, differential weighting is further validated as a sensible and necessary step when measuring poverty as people tend to assign unequal values to their human needs (Narayan 2001; Székely 2003; Halleröd 1995). However, indexes based on the human rights-based approach use equal weights given that within this framework rights are equally important (Gordon et al. 2003; Abdu and Delamonica 2017).

The applied side of differential weighting is much more complex. A problem arises due to the fact that optimal weights are unknown, and therefore any information introduced to the index is far from perfect. Hence, different approaches are proposed in the literature to set the weighting scheme of an index such as the use value or normative judgements, reliance on empirical studies about the importance of human needs (based on what people think about their needs), or inferential analyses using survey data (Decancq and Lugo 2013). Regardless of the approach, the presumption is that differential weights are a way to improve measurement and increase the likelihood of adequate discrimination between the poor and the not-poor group.

Debates surrounding differential weighting in poverty research have overlooked what has been put forward from the perspective of measurement theory (Gulliksen 1950; Retzlaff et al. 1990; Streiner et al. 2015). From the perspective of classical test theory (CTT) and latent variable measurement theories, such as item response theory (IRT) or, more generally, the latent variable approach, measurement is affected by different sources of systematic and random error and therefore indexes need to be carefully assessed under a scientific (falsifiable) framework (Cudeck and MacCallum 2012; Kvalheim 2012; Brennan 2006). The answer to the question regarding differential weights, from the perspective of measurement theory, entails caring more about reliability than about weighting (Streiner et al. 2015; Thorndike and Hagen 1969). Reliability, in that is concerned with the consistency of an index across samples, leads to robust population rankings based on a score derived from an additive combination of the observed indicators. Measurement theories, albeit conceiving reliability in different ways, predict a positive association between reliability and the probability of correct classification in a sample. From theory it is known that a reliable index is a self-weighting index, and therefore high reliability would be sufficient for a good differentiation between subjects. Differential weighting is in contention reliability in that sub-optimal weights will add unnecessary noise and contribute very little to correct classification of the poor relative to the not poor, and it would only be advisable under low reliability circumstances as an attempt to correct information provided by unreliable indicators.

In poverty research, although there have been some contemporary poverty indexes that include the assessment of reliability and apply equal weights (Nandy and Pomati 2015;

Guio et al. 2016, 2017), several indices based on the Alkire-Foster family of measures like the United Nations Development Programme (UNDP) Multidimensional Poverty Index (MPI) (Alkire and Foster 2011); the Integrated Poverty Measurement Method (IPMM) (Boltvinik and Hernández-Láos 2001); and variants of the MPI like the MPI-LA (Santos and Villatoro 2016), ignore reliability and focus on differential weighting.

One of the reasons reliability is not routinely incorporated in the examination of poverty indices has to do with the lack of a clear numerical illustration of the implications of low reliability for consistent population classification and the absence of clear guidelines for applied researches to judge different reliability statistics. This paper contributes to filling this gap in poverty research by showing the relationship between reliability, weighting and entropy (good classification). This task is accomplished by using a Monte Carlo study based on a hybrid model that associates two latent classes (poor and not-poor) with a range of poverty measures with different reliabilities (Muthén 2007). The paper is organised as follows. The second section provides an overview of weighting in poverty research and reviews reliability and its relationship with population classification. The third section describes the methods utilised to assess the main question of the paper, and the fourth section presents the findings. The final section discusses the results and its implications for poverty research.

2 Weighting and Reliability

This section introduces key concepts to discuss the relationship between weighting, reliability and classification. It first reviews the core literature on weighting in poverty research, and then it briefly presents the literature on both reliability and population classification.

2.1 Weighting

Concerns with weighting schemes are present in some of the seminal works on uni and multidimensional poverty measurement (Sen 1979; Atkinson 1987). Unlike other disciplines, such as psychometrics (Streiner et al. 2015, see below), this concern has remained theoretical and more recently has gained relevance from an empirical perspective (Decancq and Lugo 2013). The question of weighting in poverty research revolves around the best way to aggregate a set of (x_m) indicators that measure low standards of living, using an index $I(x)$. The question is whether indicators should reflect the uneven contribution of each x_m by applying differential weights ($w_m \neq 1$) to these indicators. The justification for differential weighting rests on conceptions about the nature of human needs and about how the interactions between these needs shape the idea of comparable living standards across populations (Sen 1999; Nussbaum 2001; Dean 2010). Approaches like unsatisfied basic needs (UBNs), interpretations of Sen's capability theory and other theories on human needs suggest that people find some needs more important or relevant than others, or, regardless of people's feelings, that these needs contribute differently to individual's standards of living. For example, from a physiological perspective, one question could be asked about whether eating three meals a day is more important than having a washing machine.

The main problem is that the optimal set of $w_m \neq 1$ is unknown, and therefore there is the risk of worsening the capacity $I(x)$ of indicating the correct prevalence rate and classifying individuals adequately. As a result, the literature proposes many ways to select and apply weights. Decancq and Lugo (2013), in this regard, produced a comprehensive review

of a number of approaches to weighting and classified the different schemes into three main categories: *data-driven, hybrid and normative*. Data-driven weights do not rely on the value of judgement, the most common example of which is weighting inversely according to the proportion of deprivation, i.e. items with low deprivation will have a higher weight (Desai and Shah 1988; Muffels 1993). Others opt for using model-driven approaches based mostly on factor loadings from a factor model (Krishnakumar and Nagar 2008). This, in theory, should result in an optimal approach, as the best set of weights is derived from the factor loadings. However, this paper will show how an unreliable measure results in low loadings and in turn in low reliability.

Normative weights, in contrast, draw on value judgements about the importance of human needs. The question is, how were these criteria constructed? Given that inter-personal preferences vary a lot, it is impossible to aggregate values in society to produce an incontestable set of weights based on the ranking of the necessities of life (Arrow 1950). Some UBN applications utilise this strategy for weighting dimensions (Boltvinik and Hernández-Láos 2001), while hybrid approaches aim at partially compensating for the shortcomings of normative and data-driven weighting schemes. By looking at both individual standards of living and individual valuations of one's well-being, hybrid weighting aims at finding middle ground between arbitrariness and data-driven approaches. Halleröd (1995), for instance, uses the proportion of the population that consider a given item as necessary (i.e. preference weighting)—this hybrid approach uses actual data to validate the use of differential weighting based on average preferences in society. In a similar fashion, Battiston et al. (2013) use the voices of the poor to specify the weights for each dimension in their index.

In comparison to other disciplines with a longer tradition in measurement, such as psychometrics (see below), the empirical assessment of the effects of weighting is fairly recent in poverty research. Pasha (2017), for instance, examined the effect of the MPI weighting scheme on the ranking of 28 countries. The main finding was that the position of the countries was very sensitive to the weighting scheme. Similarly, Ravallion (2012) examined the perfect substitutability (equal weights) of the HDI and produced an alternative approach with more lenient trade-offs between dimensions. Additionally, Guio et al. (2009) looked at the effect of preference weights on the ranking of countries, concluding that different sets of weights had limited effects on ranking but did have an effect on deprivation rates. Abdu and Delamonica (2017) showed that weighting might add unnecessary complexity and blur any understanding of child poverty on a global scale, arguing that sensible indicators and thresholds are a much better and simpler way to give an account of child poverty. This seems to place more emphasis on valid and reliable measurement than on weighting.

The study of the effects of weighting has a long track record in other disciplines that have established formal frameworks for weighting since the 1950s and assessed the effect of weights from the early 1970s onward (see for a comprehensive overview, Streiner et al. 2015). Wainer (1976) and Lei and Skinner (1980) reviewed the psychometric literature at that time and came to the conclusion that if items contributing poorly to the total score were eliminated, it would not matter a great deal if one were to apply weights. Basically, this suggests that there is little difference between using differential and non-differential weights, once under-performing indicators are dropped from the analysis. Once these variables are dropped, correlation between random versions of the weights in such circumstances was 0.97 (Streiner et al. 1981). In his classic contribution, Gulliksen (1950) proposed an equation to compute the correlation between weighted and unweighed measures. In this equation, the number of items, the average inter-item correlation and the standard deviation of the weights contributed to the high or low correlations between weighted

and unweighed measures. Using this statistic, Retzlaff et al. (1990) showed that empirical data are close to those predicted by Gulliksen (1950). Perloff and Persons (1988) established that when variables are not highly correlated, equal weights are likely to produce large biases. Unlike the assertion of Wainer (1976), equal weighting is likely to reduce the explanatory power of a measure, particularly when the measure is not homogeneous, i.e. unreliable (see below). In this case, differential weighting might help correct an index, provided the set of "unknown" weights is correct. This applied studies and measurement theory suggest that reliability plays a decisive role in population classification.

2.2 Reliability

The lesson from applied psychometric literature is that poor measurement of the latent construct will inevitably require a form of correction, and weighting may be a solution in this regard, and vice versa good measurement properties would not imply a form of correction. Reliability is a fundamental scientific concept employed to reflect on and assess the errors (data collection, sampling, researcher's value judgements) inherent in any measurement exercise (Streiner et al. 2015). The existence of many terms referring to reliability, such as internal consistency, precision and reproducibility, creates confusion with regard to the actual meaning of this concept. Reliability ρ can be best defined as homogeneity in measurement (Revelle and Zinbarg 2009; Streiner et al. 2015), i.e. the capacity of a series of observed measurements to capture consistently a certain construct, such as intelligence, depression or poverty. Therefore, even in the case of multidimensional measures, reliability is concerned with the capacity of observed indicators of capturing the higher-order latent construct, i.e. overall poverty (Revelle 1979). This, of course, should not be confused with *validity*, which is a concept concerned with the accuracy of a measure, i.e. the extent to which observed deprivations capture poverty and not another phenomenon. Yet, reliability is a necessary condition for a valid measure (Streiner et al. 2015; Brennan 2006).

Within psychometric theory, reliability is conceptualised and measured slightly differently across the main three measurement frameworks, namely classical test theory (CTT), item response theory (IRT) and generalisability theory (G-Theory) (Streiner et al. 2015; De Champlain 2010). However, they do share common ground in key aspects such as the ideas of homogeneity and replication, errors in measurement and population classification. In the following, the discussion focuses on CTT and IRT, given G-theory is an extension of CTT that focuses on assessing multiple sources of error in measurement and is therefore not central to the interests of this paper (Cronbach et al. 1963).

CTT draws on classic statistics (frequentist) and the idea that perfect measurement (e.g. the true parameter of interest) is affected by both random and systematic errors (Spearman 1904; Lord 1952; Novick and Lewis 1967). Spearman (1904) formally postulated that a latent true score is just an observed score plus an error. CTT assumes that among all the possible indexes there is, at least, an index that measures poverty perfectly, i.e. the true index. Because the true index is virtually unattainable, reliability uses the concept of attenuation as a means of reflecting the relationship between observed scores and the latent true score. The implication of a reliable score for measurement practices is that an imperfect index will be correlated highly with the true score, and because of this approximation or attenuation, according to CTT, clearer conclusions can be reached about the relative position of individuals in a sample, as the index tells apart the sub-population in the cohort (e.g. high achievers relative to low achievers, the better off relative to the worse off) (Streiner et al. 2015).

Unlike CTT, item response theory (IRT) does not assume that there is a true/perfect index (De Champlain 2010); instead, it argues that indicators are manifestations of an underlying phenomenon and that its properties are related, for example, to an individual's latent levels of ability, depression or poverty (Hambleton and Swaminathan 1991; Hambleton and Jodoin 2003). Furthermore, IRT is concerned with assessing the relationship between each indicator and the latent construct by using a series of parameters—in the case of a two-parameter model discrimination a_{xj} and severity b_{xj}, where $j = 1$, as there is only one factor/dimension and x is a binary item of an index (see below) (only difficulty (severity) in the case of one parameter) (Harris 1989). While severity captures how extreme, or not, the indicator is as a measure of latent overall poverty, discrimination indicates how well an indicator distinguishes the latent poor from the latent not-poor (Guio et al. 2016). Hence, discrimination is a key aspect of reliability, because the more and better the indicators tell subjects apart, the higher the reliability. If a series of indicators exhibits low discrimination, it would mean that between-subjects differences are not caused by the latent construct but by something else that is not part of the measure. This outcome is undesirable, though, especially given that one of the main goals in poverty measurement is to be able to say that one group is effectively worse-off than another. An indicator with a weak relationship with the latent construct will tell very little about the relative position of two or more individuals with regards overall poverty, for example.

At this point, it is important to spell out the links between IRT and factor analysis, given that they highlight the association between discrimination and reliability. The mathematical bases of the connection between IRT and factor analysis are formally stated by Muthén (2013). IRT can be seen as a special case of factor analysis, i.e. a unidimensional factor model. The latent variable is just measured by a mean a with variance ψ, and so $f = \alpha + \sqrt{(}\psi)\theta$, where θ is merely ability (a latent variable in an IRT model). The discrimination parameter is defined as follows:

$$a_{xj} = \lambda_{xj}\sqrt{\psi} \tag{1}$$

From this equation, factor loadings λ_{xj} relate to the power of the indicators (a_{xj}) in a two-parameter IRT model to tell apart individuals. This connection between λ_{xj}, a_{xj} and ψ affects the likelihood of (mis)classification has been formally established via the relationship between IRT modelling and Latent Class Analysis (LCA) and it occurs in the following way (Yamamoto 1987). Low discrimination (low factor loadings) imply a weak relationship between the factor and the observed deprivations meaning that changes in latent poverty will not necessarily lead to a change in observed deprivation. This weak association translates into low reliability as such an indicator will not add to the homogeneity of the measure in question. Consequently, the observed score under unreliability (many indicators with low discrimination) will not be useful to rank individuals as their observed deprivations are inconsistent as a whole. Therefore, reliability, in assessing the consistency of a measure, establishes how well a measure can rank or even split (i.e. discriminate) groups like countries or population latent groups (e.g. poor and not-poor). This applies to CTT as well, given that reliability aims to tell, given some variability between groups, how a particular subject differs from all other individuals in the sample (Revelle and Condon 2013, in press). Poverty research, for example, assumes the existence of two latent groups and seeks to tell dissimilar people apart, albeit in theory. This is where validity becomes relevant, as reliability ensures splitting two groups but validity guarantees that the two are the not poor and the poor: reliability is a precondition for validity (Guio et al. 2009, 2016).

One of the consequences of reliability, therefore, is the high probability of consistent classification in a sample (Streiner et al. 2015). However, reliability does not inform

us about how many people have been misclassified; instead, it indicates (using different indexes that range from 0 to 1, as illustrated in the next section) whether, for a set of indicators, the latent variable is measured consistently, which should hold across different measurements based on the same index. The consequences of reliability and the probability of misclassification have remained at the theoretical level. For many years the problem, from an analytical perspective, has been assessing how people move from one group to another, given a k threshold and different levels of reliability. The other problem is that there has been no clear way to measure misclassification (see Sect. 2). One of the earliest empirical explorations of the relationship between reliability and classification was provided by Thorndike and Hagen (1969), who illustrated the idea behind using rankings and therefore avoiding the problem of pre-setting two unknown groups. The authors relied on percentages to find changes in ranking positions and found that a reliability (measured by α, see definition below) above 0.80 is the minimum required to guarantee stability in ranking. Although they were unable to test this hypothesis further for a large number of parameters and scenarios, it nevertheless supported the idea that reliability is a necessary condition for correct population classification. The computational design to assess this idea further is explained in Sect. 2.

3 Methods: Simulation Strategy

A two-step procedure was utilised to assess the relationship between reliability and classification. First, data were generated using a factor mixture model (FMM), which is a hybrid model combining a measurement model (factor model) and mixture modelling (latent classes) (Muthén 2007). Therefore, this permits us to simulate: (1) a poverty measure comprising a series of binary indicators (with different discrimination and severity parameters, below) and (2) two latent groups—the poor and not-poor. The second step uses the generated data and computes different reliability ρ statistics (α, ω_t and ω_h) (See Sect. 3.1) and an E entropy index (i.e. a measure of correct classification, see Sect. 3.1) by fitting an LCA to each of the simulated datasets. Details of the simulation and the statistics utilised to evaluate the relationship between reliability and entropy are provided below.

The key changing parameters for these models are as follows:

- Number of j dimensions.
- Strong or weak dimensionality (λ_j). Loadings from the dimensions to the higher-order latent construct.
- Item reliability. Item loading parameters (λ_{xj}).
- Sample size n. Small ($n = 500$), medium ($n = 1000$) and large ($n = 5000$).
- Number of items x. Deprivation binary indicators. $x = 16$ and $x = 9$.

The paper aims to assess the relationship between reliability and entropy for measures that are commonly found in the poverty measurement literature, and therefore the array of possible measures might not capture other kinds of indexes used in sociology or economics. To simulate sensible poverty indexes, the simulations draw upon the structure of existing multiple-country, multidimensional poverty measures (Guio et al. 2016, 2017; Whelan et al. 2014; UNDP 2014; Santos and Villatoro 2016). The European measure, for example, produced by Guio et al. (2017), considers material and social deprivation but has weak dimensionality. This index has high reliability values $\alpha > 0.8$ and $\omega > 0.8$ and all

its discrimination parameters are > 0.9 (standardized loadings > 0.5). Other multidimensional measures, such as the Multidimensional Poverty Index (MPI) or the multidimensional indexes in Europe, have between three and five dimensions, where the specification of each dimension and their discriminant validity might have an impact on the relationship between the higher-order factor and the indicators, i.e. some dimensions and indicators are associated more strongly with the higher-order factor (e.g. overall poverty) than others. The MPI is meant to work well in countries with acute poverty, however, for both the 2011 Uganda measure and the 2013 Nigerian measure: $\alpha < 0.7$ and $\omega < 0.7$, and several items have low discriminations < 0.9. Similarly, low reliability values were found for the MPI Latin America proposed by Santos and Villatoro (2016) were for most countries α and ω_t are below 0.70 (based on own calculations, see "Appendix" section). These values were used as benchmark for the unreliable multidimensional measures.

Three main kinds of indexes were considered for j and λ: unidimensional, weak multidimensional and strong multidimensional measures. For the unidimensional measure ($j = 0$), there were only relationships between λ_x and the latent construct. For the case of strong dimensionality, ($j = 3$) λ_j were adjusted so that the dimensions accounted for more variations than the higher-order factor. The final case (weak dimensionality) consisted of $j = 3$, and λ values were specified to increase the relationship between the indicators and the higher-order factor. In all cases, λ_j and λ_{xj} were unequal. Therefore, the simulations set up a situation in which differential weighting was portrayed by the data, in order to ensure that the models tested a situation in which unequal weighting might be desirable/required, and to include the assumption that dimensions and specific material social deprivations have different levels of importance.

A total of 48 main models (measures of poverty $I(x)$ with 1000 replications (48,000 $I(i)$ s) were simulated (plus the 48 sub-models with sample size $n = 500$ and 48 sub-models with $x = 9$ for sensitivity analysis at 1000 replications each)[1]. The simulated data were saved and then an LCA was fitted for each model, to obtain the entropy E parameter and compute the mean classification error rate (see below for a definition). The measures were simulated in Mplus 7.2 (Muthén and Muthén 2012), and these datasets were then processed in R, using the "MplusAutomation" package (Hallquist and Wiley 2016). Reliability statistics were computed in R, using the "psych" package (Revelle 2014), and the LCA models were estimated in Mplus, too.

3.1 Measures of Reliability and Entropy

CTT and latent variable modelling theories have developed different statistics to measure reliability (Guttman 1945; Callender and Osburn 1979; Novick and Lewis 1967; Revelle and Zinbarg 2009). Cronbach's α has been the most widely used reliability statistic and has helped raise awareness about the importance of reliability. However, because α is based on very strict assumptions (τ equivalence and unidimensionality), its extensive use has come at the expense of mis-measuring reliability (Zinbarg et al. 2005). Modern statistical theory has advanced a great deal in assessing the properties of different reliability statistics, and it is now widely accepted that α, in most situations, will not work, because measurements

[1] Medium sample sizes are not reported herein, due to virtually no differences when compared with large samples

are rarely fully unidimensional with τ-equivalence between the items and the construct (Sijtsma 2008).

Statistics such as ω and ω_h have been proven to be the best reliability statistics for unidimensional and multidimensional measures, respectively (Zinbarg et al. 2005; Revelle and Zinbarg 2009). ω is known to be the the greatest lower bound of reliability (Revelle and Zinbarg 2009) and relies on the loadings of a factor model, thereby providing the highest possible reliability that an index can achieve via optimal weighting, i.e. no set of weights will improve reliability. Furthermore, ω_h is a measure of general factor saturation, i.e. the amount of variance attributable to one common factor. Therefore, it is the best measure of reliability when dealing with multidimensional measures, as it is typically a higher-order factor whose indicators of poverty load into dimensions λ_{xj}—and these into a general factor λ_j, i.e. overall poverty. Hence, this paper uses α, ω and ω_h as reliability measures, all of which range from 0 to 1, where values closer to 1 represent good reliability. This paper also includes ω_{gr}, which is the average ω of the total j dimensions, as it might be helpful to assess specific problems with the dimensions in the simulations.

With respect to the measurement of good or bad classification, entropy E is a widely used statistic of diversity (Hazewinkel 2001). Different statistics are available in the literature to compute entropy, but this paper uses the relative entropy parameter from Mplus, which is the most adequate, given the purposes of this paper, and draws on the formulation of Muthén and Muthén (2012). This parameter ranges from 0 to 1 and provides an estimate of how cleanly sub-populations are classified from the generated data. Values closer to 1, preferably $E \geq 0.8$ are associated with very low class-membership probabilities ($< 5\%$) and therefore with good classification in the literature (Celeux and Soromenho 1996).

4 Results

The results of the simulations are presented in the following order: unidimensional, weak multidimensional and strong multidimensional simulated measures. Table 1 provides a full list of the reliability and entropy coefficients for the simulated unidimensional measures. The first three columns correspond to the α statistic, and then the next three for ω and the remaining ones for E. The last column displays the mean classification error for each measure, i.e. the proportion of cases that should be classified in the latent poor class and belong to the not-poor group. Each row (cluster of simulations) corresponds to a model (16 in total) and its respective 1000 Monte Carlo replications. The columns provide values at the percentiles 5, 50 and 95% so that the degree of variability in the simulations can be appreciated (the plots for each set of simulations are presented in the "Appendix" section).

Consistent with the prediction of measurement theory, there is a clear positive association between ρ and E (and a classification error), indicating that the ability to measure the latent construct is connected with the capacity of an index to classify the population correctly. The models with higher discrimination parameters and consequently higher reliability and entropy. For example, measures with $\omega < 0.8$ systematically have an error above 6%, while for $\omega \simeq 0.7$ the classification error is around 10%. These results also are helpful in interpreting the discrimination coefficients of a measure. The unreliable measures

Table 1 Unidimensional models $x = 16$, $n = 5000$

Statistic	Reliability and entropy. Quantiles for each parameter									
	α			ω			E			% Error
Quantiles	Q5%	Q50%	Q95%	Q5%	Q50%	Q95%	Q5%	Q50%	Q95%	Mean
Model U1	0.16	0.19	0.22	0.17	0.19	0.22	0.19	0.23	0.37	28
Model U2	0.21	0.23	0.26	0.21	0.23	0.26	0.21	0.25	0.35	27
Model U3	0.25	0.28	0.3	0.26	0.28	0.3	0.26	0.29	0.35	25
Model U4	0.29	0.31	0.34	0.3	0.32	0.34	0.3	0.33	0.38	24
Model U5	0.34	0.36	0.38	0.35	0.37	0.39	0.34	0.37	0.4	22
Model U6	0.38	0.4	0.42	0.39	0.41	0.43	0.4	0.42	0.45	21
Model U7	0.44	0.46	0.48	0.45	0.47	0.49	0.45	0.47	0.5	20
Model U8	0.53	0.54	0.56	0.53	0.55	0.56	0.54	0.56	0.58	16
Model U9	0.59	0.6	0.62	0.6	0.61	0.62	0.61	0.63	0.64	15
Model U10	0.68	0.69	0.7	0.68	0.69	0.71	0.7	0.71	0.73	12
Model U11	0.73	0.74	0.75	0.74	0.75	0.76	0.77	0.78	0.79	9
Model U12	0.8	0.81	0.82	0.8	0.81	0.82	0.8	0.82	0.83	6
Model U13	0.88	0.88	0.89	0.88	0.89	0.89	0.89	0.89	0.9	3
Model U14	0.9	0.9	0.91	0.9	0.91	0.91	0.91	0.91	0.92	2
Model U15	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.95	2
Model U16	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	1

tended to have many indicators with ($a_{xj} < 0.9$) and therefore it is advisable to drop indicators with such a value as suggested in Guio et al. (2016).

The results for α and ω, where $\omega \geq \alpha$, are very similar and almost linear. The plots in the "Appendix" section (Figs. 1, 2) are helpful in illustrating that there is a lot more dispersion for low-reliability values. In contrast, when reliability is very high, E and ρ then have to be highly concentrated, which indicates the great degree of stability in the results when such high values are established. The simulations suggest that the minimum recommended value of entropy, $E \geq 0.80$, is associated with both α and $\omega \geq 0.80$, which means for the case of unidimensional measures that it is recommended to have ρs higher than 0.80. This is slightly higher than the widely used rule of thumb in social sciences of > 0.7 for α; however, it is important to remember that often $\omega \geq \alpha$ and it is vital to consider the gap between ω and α. The relationship between E and ρ for the set of 16 unidimensional measures holds for small samples ($n = 500$, see "Appendix" section). Furthermore, using fewer items $x = 9$ does not have a great impact on the findings.

Table 2 shows the full findings for the 16 simulated measures with strong multidimensionality, i.e. the dimensions account for more variance than the higher-order factor (relative to the weak dimensional model, below). This table presents the output for each simulated measure (rows in Table 2) with changing λ_{xj} and overall reliability, where $j = 3$. In this case, ω_h and $\omega_g r$ are reported along with α, E and the percentage of classification error. The simulations included α for two reasons: first, to give an idea about existing poverty measures that only apply α, where ω_h cannot be computed, and second, to illustrate that

Table 2 Strong multidimensional models $x = 16$, $n = 5000$

Model	Reliability and entropy. Quantiles for each parameter												
	α			ω_h			ω_{gr}			E			
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%	% Error
Model SD1	0.27	0.3	0.32	0.17	0.2	0.25	0.16	0.19	0.46	0.3	0.34	0.41	24
Model SD2	0.32	0.34	0.37	0.22	0.28	0.33	0.18	0.21	0.49	0.4	0.44	0.48	20
Model SD3	0.43	0.45	0.47	0.33	0.37	0.41	0.25	0.27	0.39	0.46	0.49	0.52	19
Model SD4	0.47	0.49	0.51	0.37	0.41	0.44	0.29	0.31	0.33	0.51	0.53	0.56	17
Model SD5	0.52	0.54	0.55	0.43	0.45	0.48	0.33	0.35	0.36	0.56	0.57	0.59	16
Model SD6	0.56	0.57	0.59	0.46	0.48	0.51	0.36	0.38	0.39	0.6	0.62	0.63	14
Model SD7	0.61	0.63	0.64	0.5	0.53	0.55	0.42	0.43	0.45	0.64	0.66	0.67	13
Model SD8	0.68	0.69	0.7	0.56	0.58	0.6	0.49	0.5	0.51	0.71	0.72	0.74	13
Model SD9	0.72	0.73	0.74	0.61	0.63	0.64	0.55	0.56	0.57	0.75	0.76	0.78	10
Model SD10	0.78	0.79	0.8	0.66	0.67	0.69	0.64	0.65	0.66	0.8	0.82	0.83	6
Model SD11	0.81	0.82	0.82	0.68	0.69	0.71	0.68	0.69	0.69	0.84	0.85	0.86	6
Model SD12	0.86	0.86	0.87	0.7	0.72	0.73	0.73	0.74	0.75	0.89	0.9	0.9	5
Model SD13	0.91	0.91	0.91	0.74	0.75	0.76	0.86	0.86	0.87	0.92	0.92	0.93	4
Model SD14	0.9	0.9	0.9	0.73	0.74	0.76	0.83	0.83	0.84	0.91	0.92	0.92	4
Model SD15	0.93	0.93	0.93	0.73	0.74	0.76	0.9	0.9	0.91	0.95	0.96	0.96	2
Model SD16	0.93	0.93	0.94	0.73	0.74	0.75	0.91	0.92	0.92	0.96	0.97	0.97	1

even in this case, a very high α would be required ≥ 8, in order to attain acceptable E values (see Figure 3, Appendix). However, there will be cases when a very high α might be a very poor indicator of ρ (Revelle and Zinbarg 2009).

For this set of multidimensional measures, ω_h is a better suited statistic to measure ρ. Hierarchical omega indicates the amount of variance accounted for by the higher-order factor. In the case of strong dimensionality, $\omega_h \geq 0.65$ seems to lead to high entropy values (with the classification error $< 6\%$). As can be appreciated, ω_h remains stationary for very high entropy (see Fig. 4 in the "Appendix" section), due to the fact that a very high ω_h would indicate unidimensionality. $\omega_g r$ indicates that the dimensions, not only the scale as a whole, require on average high reliability values ≥ 0.6. Similarly to the unidimensional measure, the results hold as well for nine-item and small samples (see "Appendix" section). However, given that there is more variability when using small samples, it is advisable to aim at higher reliability values when using very small survey data. As in the unidimensional case, differential weighting might help identification when reliability is low and there is no way to increase it.

The results for the weak multidimensional measures are shown in Table 3 (Figs. 5 and 6 in the "Appendix" section). Weak dimensionality means that the relationship between the observed deprivations and the higher-order factor is rather strong, i.e. the measure has some unidimensional features. For these kinds of measures, $E \geq 0.80$ is related to $\omega_h \geq 0.70$, which is higher than the ω_h of the strong multidimensional measures because when multidimensionality is low, the indicators need a stronger relationship with the overall construct

Table 3 Weak multidimensional models $x = 16$, $n = 5000$

Model	Reliability and entropy. Quantiles for each parameter												
	α			ω_h			ω_{gr}			E			
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%	% Error
Model WD1	0.5	0.51	0.53	0.4	0.46	0.5	0.27	0.31	0.55	0.58	0.61	0.63	14
Model WD2	0.56	0.58	0.59	0.47	0.51	0.55	0.32	0.36	0.49	0.61	0.64	0.66	13
Model WD3	0.61	0.62	0.63	0.52	0.55	0.57	0.36	0.39	0.45	0.63	0.65	0.67	12
Model WD4	0.64	0.65	0.66	0.56	0.58	0.61	0.41	0.43	0.45	0.67	0.69	0.7	12
Model WD5	0.68	0.7	0.7	0.6	0.62	0.64	0.45	0.47	0.49	0.71	0.72	0.74	12
Model WD6	0.71	0.72	0.73	0.63	0.65	0.67	0.49	0.5	0.51	0.74	0.75	0.76	12
Model WD7	0.76	0.76	0.77	0.67	0.69	0.7	0.55	0.56	0.57	0.77	0.78	0.8	10
Model WD8	0.8	0.81	0.81	0.71	0.73	0.74	0.61	0.62	0.63	0.82	0.83	0.84	6
Model WD9	0.83	0.84	0.84	0.75	0.76	0.77	0.66	0.67	0.68	0.85	0.86	0.87	5
Model WD10	0.87	0.87	0.88	0.78	0.79	0.8	0.73	0.74	0.75	0.89	0.89	0.9	4
Model WD11	0.89	0.89	0.89	0.8	0.81	0.82	0.76	0.77	0.77	0.9	0.91	0.92	4
Model WD12	0.92	0.92	0.92	0.82	0.83	0.84	0.81	0.82	0.82	0.93	0.93	0.94	4
Model WD13	0.94	0.95	0.95	0.84	0.85	0.86	0.9	0.91	0.91	0.95	0.95	0.96	3
Model WD14	0.94	0.94	0.94	0.84	0.85	0.85	0.88	0.88	0.89	0.94	0.95	0.95	3
Model WD15	0.96	0.96	0.96	0.84	0.84	0.85	0.93	0.93	0.93	0.96	0.97	0.97	2
Model WD16	0.96	0.96	0.96	0.83	0.84	0.85	0.94	0.94	0.94	0.97	0.98	0.98	1

to increase reliability, so that this in turn compensates for the seemingly weak association with the dimensions.

Table 3 shows that entropy is systematically above 0.80 when $\omega_{gr} \geq 0.6$. The low value of the ω_{gr} is due to the fact that in the case of weak dimensionality, the reliability of the dimensions matters a little less, since reliability with regard to the latent construct (higher-order factor) gains importance. For small samples, the simulations have more variability and $\omega_h \geq 0.65$ seems to be enough to achieve the minimum recommended value of entropy. Similarly, for the nine-item index, $\omega_h \geq 0.65$ is systematically associated with high Es (see "Appendix" section). The minimum value of entropy for the most unreliable measures seems a bit high at 0.60 in comparison to previous estimates, one reason for which is that for this measure the minimum ω_h was 0.40, which is higher than in the previous example, due to the difficulty inherent in fitting a model with weak dimensionality and very low factor loadings. Such a model would mean that neither dimensionality nor reliability is present in the model and it would be difficult to identify a measure of this kind in the contemporary literature.

These simulations suggest that in order to increase E, the differential weighting scheme should pay more attention to the weights of the items and a bit less to the weights of the dimensions, as they are not strongly associated with overall poverty. Measures with low discriminant validity are more likely to experience this kind of behaviour, for example when the majority of the indicators load into a couple of dimensions and these in turn are highly correlated. In this situation, it is likely that the measure is capturing only one aspect of poverty, and therefore using one dimension should lead to a better representation of the data.

5 Conclusion and Discussion

This paper contributes to the existing literature by providing an numerical side to the claim that reliability increases the probability of correct population classification in the context of unidimensional and multidimensional measurement. The consequence of this claim is that reliability guarantees a self-weighting index, and therefore differential weighting would add little, even if the weights were close to perfection, to the likelihood of correct classification of the poor and the not-poor groups.

Previous studies in the field of psychometrics have attempted to explore the relationship between reliability, classification and consistent measurement when using different weights (Gulliksen 1950; Thorndike and Hagen 1969; Perloff and Persons 1988). This paper took this idea further and investigates from an numerical perspective the relationship between dimensionality, reliability and entropy, using a more adequate and powerful approach. Using a Monte Carlo study based on factor mixture modelling, the study simulated 144 main types of poverty measures (with 1000 replications), specifying different dimensional structures (unidimensional, multidimensional and weak multidimensional) and diverse relationships between binary indicators and latent constructs. This therefore allowed for generalising the results to a wider set of conditions.

The results back up the idea that reliability increases the likelihood of correctly classifying units in a sample (Streiner et al. 2015). Therefore, the findings suggest that reliability is a condition necessary to produce a poverty index that correctly classifies the poor and the not-poor, provided the index is valid. This paper suggests that for unidimensional measures, α and ω_t above 0.80 are very likely to produce high entropy (> 0.8), leading to a much higher likelihood of good distinction between the poor and the not-poor where the classification error is $< 5\%$. For multidimensional measures with well-defined dimensions, the recommendation, following the contemporary literature, is to avoid α and use ω_h (Revelle and Zinbarg 2009; Zinbarg et al. 2005). ω_h values above 0.65 (and $\omega_t > 0.8$) are the very minimum expected for a poverty index that differentiates between the poor and the not-poor (error $< 5\%$). For a strong multidimensional measure, it is advisable to compute ω_{gr} to identify dimensions with measurement problems. Low reliability should be a major concern and it should be routinely examined in poverty measurement. The findings suggest that α and ω_t below 0.80 are likely to result in classification error above 10% (i.e. 10% of the poor will be classified as not poor). Similarly, for multidimensional measures, ω_h values below 0.65 (and $\omega_t < 0.8$) result in classification error above 10%.

The consequence of these findings is that, under reliability, differential weighting would contribute little to improving the classification or ordering of individuals. Equal weighting will work in circumstances where high reliability is guaranteed, which is consistent with previous studies on the effects of weighting on maximising the predictive accuracy of a set of weights. For example, Wainer (1976) suggest that differential weighting is unnecessary when dropping items making little contribution to the score (i.e. items with low reliability within contemporary measurement frameworks). Lei and Skinner (1980) and Streiner et al. (1981) show an empirical application of this principle. Similarly, Guio et al. (2009) found little effect of weights, which seems to be related to the use of a more reliable measure.

In poverty research, it has been shown that weighting makes a difference (Pasha 2017; Abdu and Delamonica 2017; Ravallion 2012), albeit this has been shown to have unreliable measures such as the Global MPI and the Human Development Index (HDI). Furthermore,

in terms of applied research, drawing upon the findings of these simulations, the reliability values of both the MPI and the MPI-LA raise serious concerns about the misclassification of the poor (for Nigeria and Uganda would be above 10%; likewise for Chile, Brazil, Uruguay, Argentina, Bolivia, Mexico) as the values reported in this study are a lower bound, i.e. do not consider the further noise introduced by weights and the potentially misspecified dimensional structure. Therefore, this reasserts the extent of the problem in real application and that unreliability could be not only corrected, but also affected by differential weighting.

When high reliability is not feasible, due to data limitations (i.e. not enough variables or limitations to compare countries across a range of indicators), differential weighting has the potential to be helpful. In these circumstances, though, equal weighting is likely to produce severe bias (Perloff and Persons 1988), which, according to this study, might happen in the presence of low reliability. However, working with an unreliable index carries other problems that weighting itself is unlikely to solve. Reliability is a condition for validity, and therefore having an unreliable index will reveal little about whether poverty is actually measured by the indicators comprising an index (Guio et al. 2016). If using differential weights is unavoidable, then it is strongly recommended to compute (Gulliksen 1950)'s correlation measure and see which factor is contributing to the high/low association between pairs of indexes.

The findings of this paper have important implications for current practices in poverty measurement, particularly when assessing the properties of an index. The reliability values found in this paper validate the thresholds for the standardised discrimination parameters (e.g. > 0.4) used by (Guio et al. 2016, 2012; Nandy and Pomati 2015). The emerging practice of analysing reliability in poverty research certainly should continue and be encouraged in poverty research (Saunders and Naidoo 2009; Fusco and Dickens 2008; Nolan and Whelan 2010; Szeles and Fusco 2013). However, it is important to stress that although reliability increases the likelihood of correct classification, there might be other reasons for using weights, such as placing emphasis on the most extreme forms of poverty and making comparisons across years. This latter topic, nonetheless, remains virtually unexplored by the empirical literature.

As a final note, the simulated conditions of this paper were expected to cover the most common situations found in poverty research, namely unidimensional deprivation, weak multidimensional and strong multidimensional measures. Nevertheless, it is important to continue working on the effects of different numbers of items, unbalanced dimensions (i.e. dimensions with many indicators, and others with very few), extreme poverty rates, etc. Furthermore, it would be important to cross-validate the findings of this study, by computing (Gulliksen 1950)'s statistic, so that we have a better idea, aside from entropy, of what happens with the weights.

Appendix

See Table 4 and Figs. 1, 2, 3, 4, 5 and 6.

Table 4 Reliability statistics. MPI-LA. *Source*: Own estimates

Country	Year	α	ω
Argentina	2005	0.63	0.71
	2012	0.51	0.6
Bolivia	2003	0.64	0.68
	2012	0.65	0.76
Brazil	2005	0.52	0.62
	2012	0.45	0.57
Chile	2003	0.46	0.58
	2011	0.27	0.33
Mexico	2004	0.75	0.81
	2012	0.64	0.69
Uruguay	2005	0.54	0.67
	2012	0.43	0.54
Nigeria (MPI)	2013	0.67	0.71
Uganda (MPI)	2011	0.61	0.68

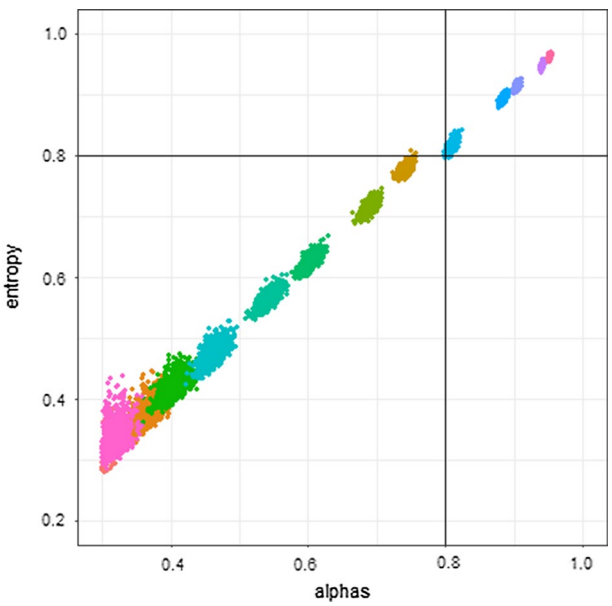

Fig. 1 Unidimensional models α and E

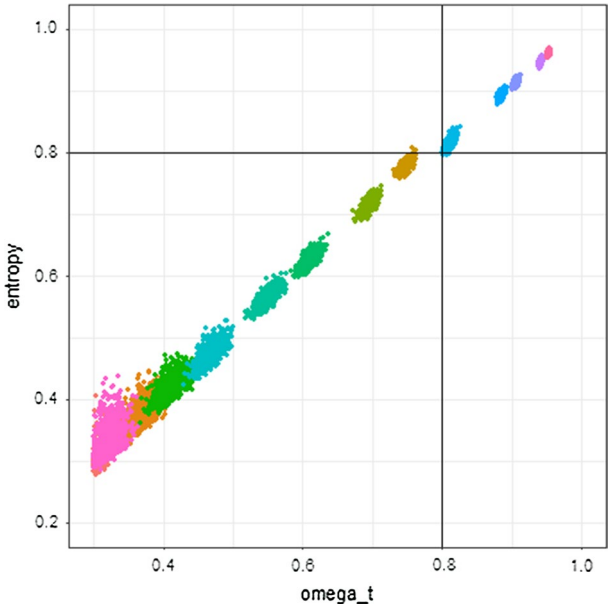

Fig. 2 Unidimensional models ω and E

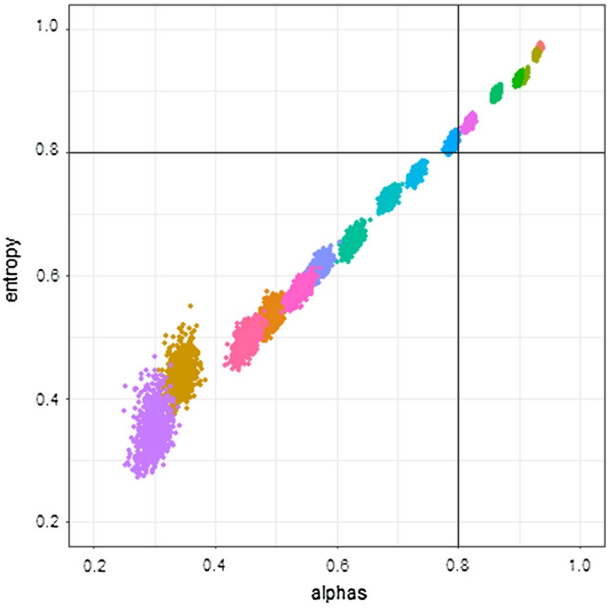

Fig. 3 Strong multidimensional models α and E

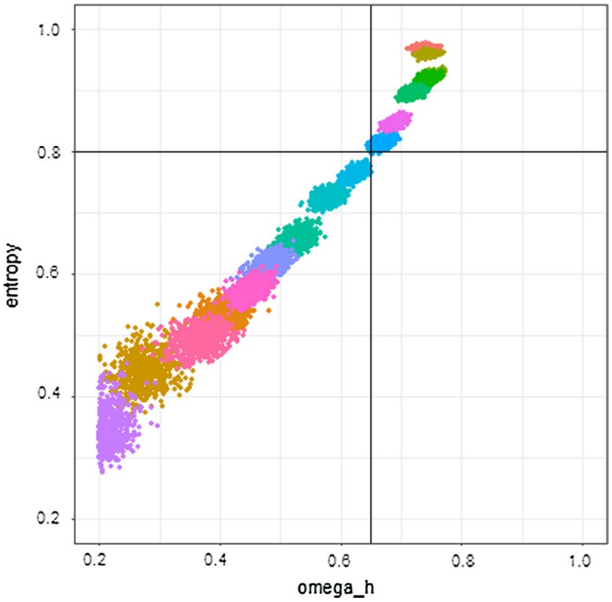

Fig. 4 Strong multidimensional models ω_h and E

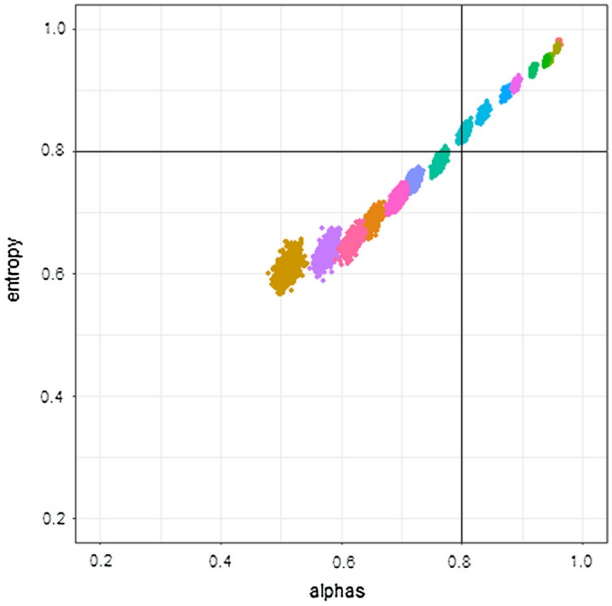

Fig. 5 Weak multidimensional models α and E

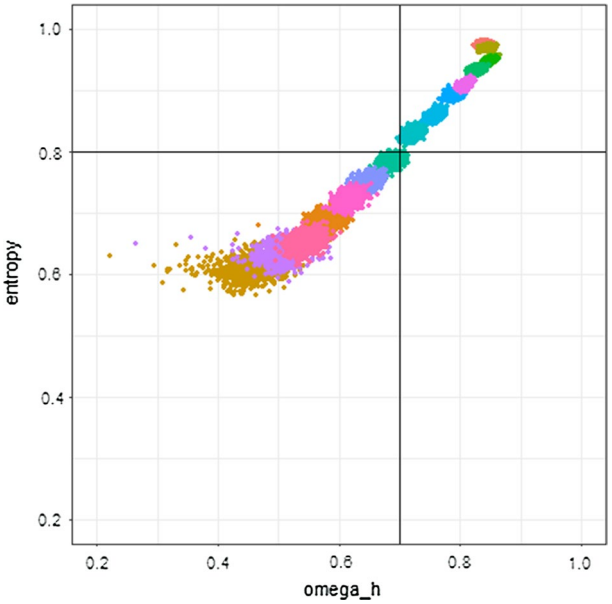

Fig. 6 Weak multidimensional models ω_h and E

	Unidimensional model (16-item). $n = 500$								
Model	α			ω			E		
Quantile	5%	50%	95%	5%.1	50%.1	95%.1	5%.2	50%.2	95%.2
stats_2	0.25	0.33	0.39	0.26	0.34	0.39	0.35	0.43	0.73
stats_16	0.21	0.29	0.35	0.23	0.31	0.36	0.33	0.44	0.80
stats_15	0.30	0.37	0.43	0.32	0.38	0.43	0.38	0.45	0.58
stats_12	0.35	0.41	0.46	0.36	0.42	0.47	0.42	0.48	0.57
stats_13	0.16	0.24	0.31	0.18	0.27	0.33	0.31	0.50	0.91
stats_8	0.40	0.47	0.51	0.42	0.48	0.52	0.46	0.52	0.60
stats_3	0.11	0.19	0.28	0.17	0.24	0.30	0.30	0.54	0.96
stats_9	0.49	0.55	0.59	0.50	0.56	0.59	0.53	0.59	0.65
stats_10	0.56	0.61	0.65	0.57	0.62	0.65	0.60	0.65	0.70
stats_11	0.66	0.69	0.72	0.66	0.70	0.73	0.69	0.73	0.78
stats_14	0.72	0.74	0.77	0.73	0.75	0.78	0.76	0.79	0.83
stats_7	0.79	0.81	0.83	0.79	0.81	0.83	0.79	0.82	0.86
stats_6	0.87	0.88	0.90	0.87	0.89	0.90	0.88	0.90	0.92

	Unidimensional model (16-item). $n = 500$								
Model	α			ω			E		
Quantile	5%	50%	95%	5%.1	50%.1	95%.1	5%.2	50%.2	95%.2
stats_5	0.89	0.91	0.91	0.90	0.91	0.92	0.90	0.92	0.94
stats_4	0.93	0.94	0.95	0.94	0.94	0.95	0.93	0.95	0.97
stats_1	0.95	0.95	0.96	0.95	0.95	0.96	0.95	0.97	0.98

	Unidimensional model (nine-item). $n = 5000$								
Model	α			ω			E		
Quantile	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_15	0.17	0.27	0.33	0.21	0.29	0.34	0.27	0.40	0.67
stats_12	0.18	0.28	0.34	0.23	0.31	0.35	0.28	0.40	0.63
stats_8	0.24	0.33	0.39	0.28	0.35	0.40	0.35	0.42	0.53
stats_2	0.11	0.20	0.27	0.16	0.23	0.28	0.25	0.47	0.88
stats_9	0.35	0.43	0.48	0.38	0.44	0.49	0.43	0.50	0.57
stats_13	0.06	0.17	0.24	0.14	0.20	0.25	0.27	0.55	0.96
stats_16	0.07	0.17	0.25	0.14	0.21	0.26	0.26	0.57	0.92
stats_10	0.46	0.51	0.55	0.48	0.53	0.56	0.50	0.57	0.63
stats_3	0.04	0.15	0.23	0.14	0.19	0.24	0.25	0.60	0.93
stats_11	0.58	0.62	0.65	0.59	0.63	0.66	0.63	0.67	0.71
stats_7	0.64	0.67	0.70	0.64	0.67	0.71	0.63	0.69	0.73
stats_14	0.64	0.67	0.69	0.65	0.68	0.70	0.69	0.73	0.77
stats_6	0.81	0.83	0.84	0.81	0.83	0.84	0.82	0.85	0.87
stats_5	0.86	0.87	0.88	0.86	0.87	0.88	0.86	0.89	0.90
stats_4	0.89	0.90	0.91	0.89	0.90	0.91	0.90	0.92	0.93
stats_1	0.90	0.91	0.92	0.91	0.92	0.92	0.93	0.94	0.96

	Strong multidimensional model (16-item). $n = 500$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_13	0.23	0.30	0.38	0.03	0.14	0.24	0.14	0.25	0.44	0.37	0.48	0.85
stats_3	0.26	0.35	0.42	0.08	0.19	0.32	0.19	0.27	0.47	0.41	0.50	0.75
stats_16	0.39	0.45	0.51	0.15	0.25	0.37	0.25	0.33	0.52	0.47	0.53	0.62
stats_2	0.44	0.49	0.55	0.20	0.31	0.42	0.27	0.35	0.59	0.51	0.57	0.65
stats_15	0.49	0.54	0.59	0.26	0.37	0.49	0.29	0.37	0.45	0.55	0.60	0.65
stats_12	0.53	0.57	0.61	0.29	0.41	0.53	0.33	0.40	0.59	0.58	0.64	0.68
stats_8	0.59	0.62	0.66	0.35	0.45	0.55	0.38	0.45	0.65	0.63	0.68	0.73
stats_9	0.66	0.68	0.71	0.45	0.52	0.61	0.46	0.50	0.56	0.69	0.74	0.77
stats_10	0.70	0.73	0.76	0.51	0.57	0.65	0.52	0.56	0.60	0.74	0.77	0.81
stats_11	0.77	0.79	0.81	0.58	0.63	0.68	0.62	0.65	0.67	0.79	0.82	0.86
stats_14	0.80	0.82	0.84	0.62	0.67	0.71	0.66	0.69	0.71	0.83	0.86	0.88
stats_7	0.85	0.86	0.87	0.66	0.70	0.74	0.70	0.73	0.76	0.88	0.90	0.92
stats_6	0.89	0.90	0.91	0.69	0.73	0.77	0.82	0.83	0.84	0.90	0.92	0.94

	Strong multidimensional model (16-item). $n = 500$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_5	0.90	0.91	0.92	0.70	0.74	0.78	0.85	0.86	0.87	0.91	0.93	0.95
stats_4	0.92	0.93	0.93	0.70	0.74	0.77	0.89	0.90	0.91	0.95	0.96	0.97
stats_1	0.93	0.93	0.94	0.69	0.73	0.77	0.91	0.92	0.92	0.96	0.97	0.98
Source	5%	50%	95%	5%.1	50%.1	95%.1	5%.2	50%.2	95%.2	5%.3	50%.3	95%.3
	Strong multidimensional model (nine-item). $n = 5000$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_16	0.21	0.29	0.37	0.09	0.19	0.27	0.19	0.38	0.61	0.34	0.41	0.74
stats_13	0.12	0.21	0.29	0.06	0.15	0.24	0.15	0.31	0.58	0.27	0.44	0.78
stats_2	0.23	0.32	0.39	0.08	0.19	0.32	0.20	0.38	0.60	0.35	0.46	0.61
stats_3	0.19	0.28	0.35	0.09	0.17	0.28	0.18	0.38	0.56	0.34	0.47	0.72
stats_15	0.33	0.39	0.46	0.18	0.27	0.37	0.27	0.40	0.56	0.41	0.49	0.63
stats_8	0.40	0.46	0.51	0.21	0.29	0.40	0.31	0.43	0.61	0.43	0.51	0.62
stats_12	0.36	0.43	0.48	0.16	0.29	0.41	0.29	0.42	0.65	0.43	0.52	0.63
stats_9	0.49	0.54	0.59	0.27	0.40	0.50	0.35	0.44	0.65	0.51	0.59	0.68
stats_10	0.61	0.62	0.63	0.50	0.53	0.55	0.45	0.47	0.48	0.66	0.68	0.70
stats_11	0.66	0.69	0.72	0.47	0.53	0.60	0.53	0.58	0.73	0.71	0.76	0.82
stats_14	0.69	0.72	0.75	0.51	0.57	0.62	0.57	0.62	0.67	0.74	0.79	0.83
stats_7	0.72	0.75	0.78	0.56	0.61	0.66	0.61	0.63	0.67	0.77	0.81	0.84
stats_4	0.85	0.87	0.88	0.66	0.71	0.74	0.85	0.86	0.87	0.86	0.89	0.91
stats_1	0.86	0.88	0.89	0.66	0.70	0.74	0.87	0.88	0.89	0.87	0.90	0.92
stats_6	0.81	0.83	0.84	0.63	0.68	0.72	0.74	0.76	0.78	0.88	0.90	0.92
stats_5	0.83	0.85	0.87	0.65	0.70	0.73	0.80	0.82	0.83	0.88	0.91	0.93

	Weak multidimensional model (16-item). $n = 500$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_3	0.46	0.52	0.57	0.21	0.36	0.53	0.28	0.36	0.50	0.58	0.64	0.74
stats_13	0.53	0.58	0.62	0.21	0.41	0.54	0.30	0.40	0.59	0.60	0.66	0.72
stats_16	0.58	0.62	0.66	0.32	0.45	0.58	0.37	0.43	0.64	0.62	0.67	0.73
stats_2	0.62	0.65	0.69	0.38	0.51	0.63	0.40	0.45	0.65	0.66	0.71	0.76
stats_15	0.66	0.69	0.73	0.46	0.56	0.66	0.41	0.48	0.69	0.70	0.74	0.78
stats_12	0.69	0.72	0.75	0.47	0.58	0.68	0.43	0.52	0.58	0.72	0.76	0.80
stats_8	0.74	0.76	0.79	0.56	0.63	0.70	0.49	0.57	0.68	0.76	0.79	0.84
stats_9	0.79	0.80	0.82	0.64	0.68	0.72	0.57	0.63	0.68	0.80	0.84	0.87

	Weak multidimensional model (16-item). $n = 500$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_10	0.82	0.84	0.85	0.67	0.71	0.75	0.63	0.67	0.70	0.84	0.87	0.89
stats_11	0.86	0.87	0.89	0.71	0.75	0.79	0.71	0.74	0.76	0.87	0.90	0.92
stats_14	0.88	0.89	0.90	0.74	0.78	0.81	0.74	0.76	0.79	0.90	0.92	0.94
stats_7	0.91	0.92	0.93	0.78	0.81	0.84	0.78	0.81	0.83	0.92	0.94	0.96
stats_6	0.93	0.94	0.95	0.81	0.83	0.86	0.87	0.88	0.89	0.94	0.95	0.97
stats_5	0.94	0.95	0.95	0.81	0.84	0.86	0.90	0.91	0.91	0.94	0.96	0.97
stats_4	0.95	0.96	0.96	0.81	0.84	0.86	0.93	0.93	0.94	0.96	0.97	0.98
stats_1	0.96	0.96	0.97	0.81	0.83	0.86	0.94	0.94	0.95	0.97	0.98	0.99
	Weak multidimensional model (nine-item). $n = 5000$											
Model	α			ω_h			ω_{gr}			E		
source	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
stats_16	0.40	0.47	0.51	0.20	0.33	0.42	0.29	0.45	0.59	0.47	0.53	0.61
stats_3	0.37	0.44	0.48	0.19	0.34	0.44	0.30	0.42	0.61	0.48	0.55	0.67
stats_2	0.43	0.50	0.53	0.23	0.37	0.47	0.32	0.47	0.63	0.50	0.57	0.65
stats_13	0.41	0.47	0.52	0.22	0.35	0.46	0.29	0.47	0.65	0.48	0.57	0.65
stats_15	0.53	0.58	0.61	0.33	0.46	0.55	0.37	0.48	0.71	0.57	0.63	0.68
stats_12	0.55	0.60	0.63	0.37	0.48	0.56	0.40	0.49	0.66	0.60	0.65	0.70
stats_8	0.60	0.64	0.67	0.42	0.51	0.57	0.41	0.50	0.68	0.61	0.66	0.71
stats_9	0.68	0.71	0.73	0.47	0.59	0.66	0.49	0.56	0.73	0.69	0.74	0.77
stats_10	0.74	0.76	0.78	0.59	0.65	0.70	0.57	0.62	0.76	0.78	0.82	0.84
stats_11	0.79	0.82	0.83	0.65	0.70	0.74	0.65	0.69	0.74	0.82	0.86	0.88
stats_14	0.81	0.83	0.85	0.67	0.72	0.75	0.67	0.72	0.75	0.84	0.87	0.89
stats_7	0.83	0.85	0.86	0.71	0.76	0.78	0.71	0.73	0.75	0.86	0.88	0.90
stats_1	0.92	0.93	0.94	0.80	0.83	0.85	0.90	0.91	0.92	0.92	0.94	0.95
stats_4	0.91	0.93	0.93	0.80	0.82	0.84	0.89	0.90	0.91	0.92	0.94	0.95
stats_5	0.90	0.92	0.92	0.78	0.82	0.84	0.85	0.87	0.88	0.92	0.94	0.96
stats_6	0.89	0.90	0.91	0.76	0.80	0.82	0.81	0.83	0.85	0.91	0.94	0.95

References

Abdu, M., & Delamonica, E. (2017). Multidimensional child poverty: From complex weighting to simple representation. *Social Indicators Research*, *136*, 881.

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

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, *58*(4), 328–346.

Atkinson, A. B. (1987). On the measurement of poverty. *Econometrica*, *55*(4), 749–764.

Battiston, D., Cruces, G., Lopez-Calva, L. F., Lugo, M. A., & Santos, M. E. (2013). Income and beyond: Multidimensional poverty in six latin Américan countries. *Social Indicators Research*, *112*(2), 291–314.

Boltvinik, J., & Hernández-Láos, H. (2001). *Pobreza y distribución del ingreso en México*. Siglo XXI Editores.

Brennan, R. L. (2006). *Educational measurement. ACE/Praeger Series on Higher Education*. ERIC.

Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, guttman's lambda-2, and msplit maximized split-half reliability estimates. *Journal of Educational Measurement*, *16*(2), 89–99.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195–212.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

Cudeck, R., & MacCallum, R. C. (2012). *Factor analysis at 100: Historical developments and future directions*. London: Routledge.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117.

Dean, H. (2010). *Understanding human need*. Bristol: The Policy Press, University of Bristol.

Decancq, K., & Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, *32*(1), 7–34.

Desai, M., & Shah, A. (1988). An econometric approach to the measurement of poverty. *Oxford Economic Papers*, *40*(3), 505–522.

Fusco, A., & Dickens, P. (2008). The Rasch model and multidimensional poverty measurement. In N. Kakwani & J. Silber (Eds.), *Quantitative approaches to multidimensional poverty measurement*. Basingstoke: Palgrave-MacMillan.

Gordon, D. (2006). The concept and measurement of poverty. In C. Pantazis, D. Gordon, & R. Levitas (Eds.), *Poverty and social exclusion in Birtain: The Milenium survey* (pp. 29–69). Bristol: Bristol Policy Press.

Gordon, D., Nandy, S., Pantazis, C., Pemberton, S., & Townsend, P. (2003). *Child poverty in the developing world*. Bristol: The Policy Press, University of Bristol.

Guio, A., Fusco, A., & Marlier, E. (2009). A European union approach to material deprivation using EU-silc and eurobarometer data. Technical report, international networks for studies in technology, environment, alternatives and development.

Guio, A., Gordon, D., & Marlier, E. (2012). Measuring material deprivation in the EU: Indicators for the whole population and child-specific indicators. Technical report, EUROSTAT.

Guio, A.-C., Gordon, D., Marlier, E., Najera, H., & Pomati, M. (2017). Towards an EU measure of child deprivation. *Child Indicators Research*, *11*, 835.

Guio, A.-C., Gordon, D., Najera, H., & Pomati, M. (2017). Revising the EU material deprivation variables. Technical report, EUROSTAT.

Guio, A.-C., Marlier, E., Gordon, D., Fahmy, E., Nandy, S., & Pomati, M. (2016). Improving the measurement of material deprivation at the european union level. *Journal of European Social Policy*, *26*(3), 219–333.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, *10*(4), 255–282.

Halleröd, B. (1995). The truly poor: Direct and indirect consensual measurement of poverty in Sweden. *Journal of European Social Policy*, *5*(2), 111–129.

Hallquist, M., & Wiley, J. (2016). Mplusautomation. Technical report.

Hambleton, R., & Jodoin, M. (2003). Item response theory: Models and features. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 510–515). London: SAGE Publications Ltd.

Hambleton, R., & Swaminathan, J. (1991). *Fundamental of item response theory*. London: Sage Publications Inc.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter irt models. *Educational Measurement: Issues and Practice*, *8*(1), 35–41.

Hazewinkel, M. (2001). Entropy. *Encyclopedia of Mathematics*, *1*, 1.

Kakwani, N., & Silber, J. (2008). *Quantitative approaches to multidimensional poverty measurement*. Berlin: Springer.

Krishnakumar, J., & Nagar, A. (2008). On exact statistical properties of multidimensional indices based on principal components, factor analysis, mimic and structural equation models. *Social Indicators Research*, *86*(3), 481–496.

Kvalheim, O. M. (2012). History, philosophy and mathematical basis of the latent variable approach: From a peculiarity in psychology to a general method for analysis of multivariate data. *Journal of Chemometrics*, *26*(6), 210–217.

Lei, H., & Skinner, H. A. (1980). A psychometric study of life events and social readjustment. *Journal of Psychosomatic Research*, *24*(2), 57–65.
Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, *7*(1), 84.
Muffels, R. (1993). Deprivation standards and style of living indices. *The European Face of Social Security*, *1*, 43–59.
Muthén, B. (2007). Laten variable hybrids. Overview of old and new models. In G. Hancock & K. Samuelsen (Eds.), *Advances in latent variable mixture models*. Charlotte: Information Age Publishing.
Muthén, B. (2013). Irt in mplus. Technical report, Mplus.
Muthén, L., & Muthén, B. (2012). Mplus user's guide (7th ed.). Mplus.
Nandy, S., & Pomati, M. (2015). Applying the consensual method of estimating poverty in a low income african setting. *Social Indicators Research*, *124*(3), 693–726.
Narayan, D. (2001). *Voices of the poor. Faith in development: Partnership between the world bank and the churches in Africa* (pp. 39–50). Oxford: Regnum Books.
Nolan, B., & Whelan, C. T. (2010). Using non-monetary deprivation indicators to analyze poverty and social exclusion: Lessons from Europe? *Journal of Policy Analysis and Management*, *29*(2), 305–325.
Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1–13.
Nussbaum, M. C. (2001). *Women and human development: The capabilities approach*. Cambridge: Cambridge University Press.
Pasha, A. (2017). Regional perspectives on the multidimensional poverty index. *World Development*, *94*, 268–285.
Perloff, J. M., & Persons, J. B. (1988). Biases resulting from the use of indexes: An application to attributional style and depression. *Psychological Bulletin*, *103*(1), 95.
Ravallion, M. (2012). Troubling tradeoffs in the human development index. *Journal of Development Economics*, *99*(2), 201–209.
Retzlaff, P. D., Sheehan, E. P., & Lorr, M. (1990). MCMI-II scoring: Weighted and unweighted algorithms. *Journal of Personality Assessment*, *55*(1–2), 219–223. (PMID: 2231242).
Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74.
Revelle, W. (2014). Psych. R package. Technical report, R software.
Revelle, W., & Condon, D. M. (2013). Reliability. In P. Irwing, T. Booth, & D. Hughes (Eds.), *Handbook of psychometric testing*. New York: Wiley-Blackwell. (in press).
Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, *74*(1), 145–154.
Santos, M. E., & Villatoro, P. (2016). A multidimensional poverty index for Latin America. *Review of Income and Wealth*, *64*, 52.
Saunders, P., & Naidoo, Y. (2009). Poverty, deprivation and consistent poverty. *Economic Record*, *85*(271), 417–432.
Sen, A. (1979). Issues in the measurement of poverty. *The Scandinavian Journal of Economics*, *158*, 285–307.
Sen, A. (1999). *Development as freedom*. Oxford: The Clarendon Press.
Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101.
Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford University Press.
Streiner, D. L., Norman, G. R., McFarlane, A. H., & Roy, R. G. (1981). Quality of life events and their relationship to strain. *Schizophrenia Bulletin*, *7*(1), 34–42.
Székely, M. (2003). Lo que dicen los pobres. *Cuadernos de Desarrollo Humano*, *13*, 1.
Szeles, M., & Fusco, A. (2013). Item response theory and the measurement of deprivation: Evidence from luxembourg data. *Quality and Quantity*, *47*(3), 1545–1560.
Thorndike, R., & Hagen, E. (1969). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart and Winston.
Townsend, P. (1979). *Poverty in the United Kingdom: A survey of household resources and standards of living*. Oakland: University of California.
UNDP (2014). Multidimensional Poverty Index (MPI). Technical report, UNDP.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213.

Whelan, C. T., Nolan, B., & Maitre, B. (2014). Multidimensional poverty measurement in europe: An application of the adjusted headcount approach. *Journal of European Social Policy*, *24*(2), 183–197.

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, revelle's β, and mcdonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.