University of BRISTOL

Universidad Nacional Autónoma de México
UNAM

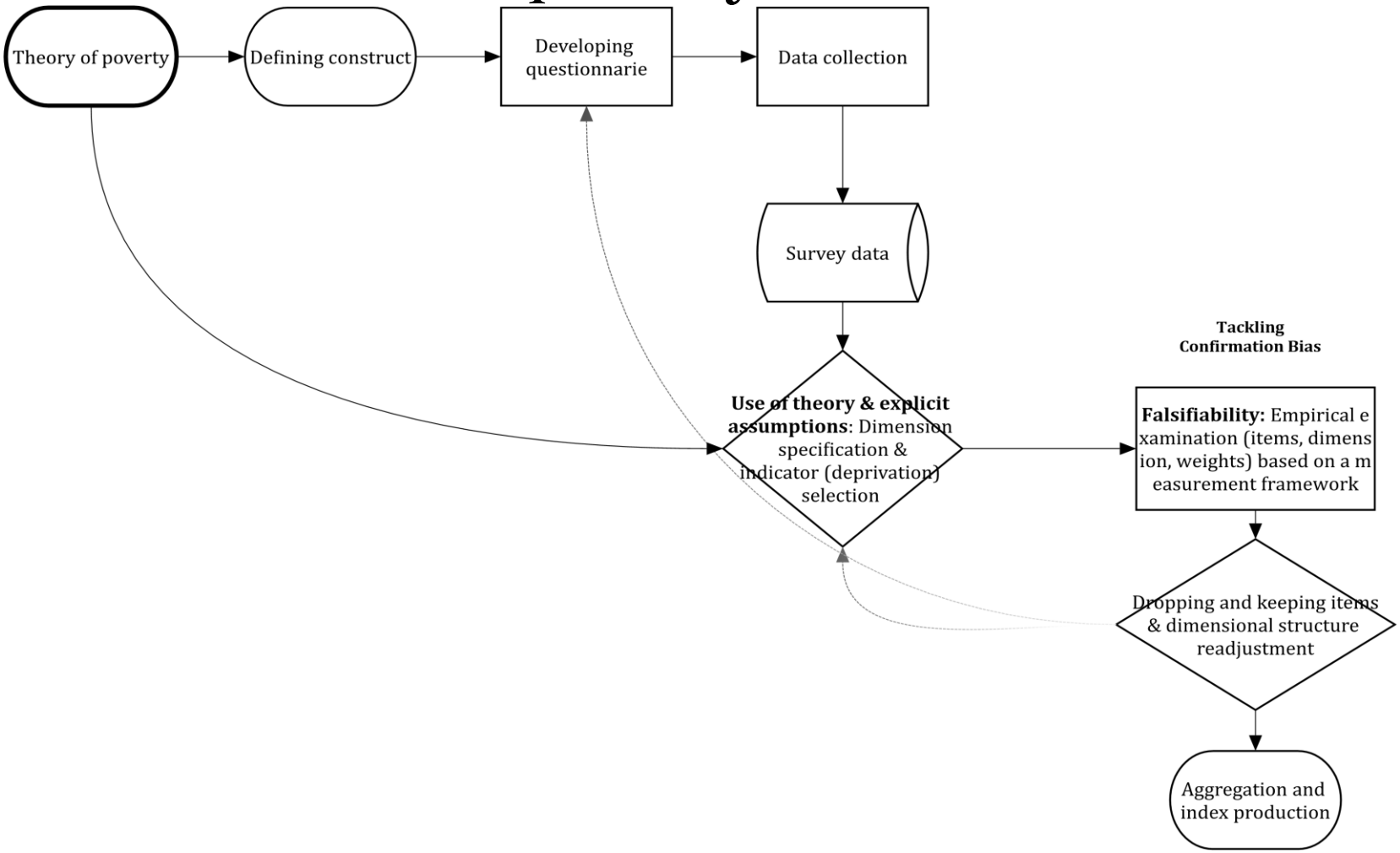
Scientific principles in poverty meaurement: Reliability and validity

Dr. Hector Najera

Outline

- Scientific measurement and poverty research
- How and why do we use scientific measurement standards in poverty research
- Two key principles in scientific measurement:
  - Reliability
  - Validity
- We will see both the theory and implementation of these two principles

Ideal workflow in poverty measurement



The ideal is not often feasible or pursued

One of the often-overlooked features in poverty measurement is that fact that researchers raise several assumptions about the nature of poverty and never work with the full set of information (perfect data) to produce a measure

- We assume there are j dimensions
- We assume that these include x indicators

Scientific measurement aims to incorporate a framework to falsify researcher's assumptions, i.e. a set of outcome indicators leads to a **good** measure of poverty. (We will define good formally)

A set of principles that make poverty measures reproducible, i.e. two researchers should arrive to the same set of indicators and dimensions

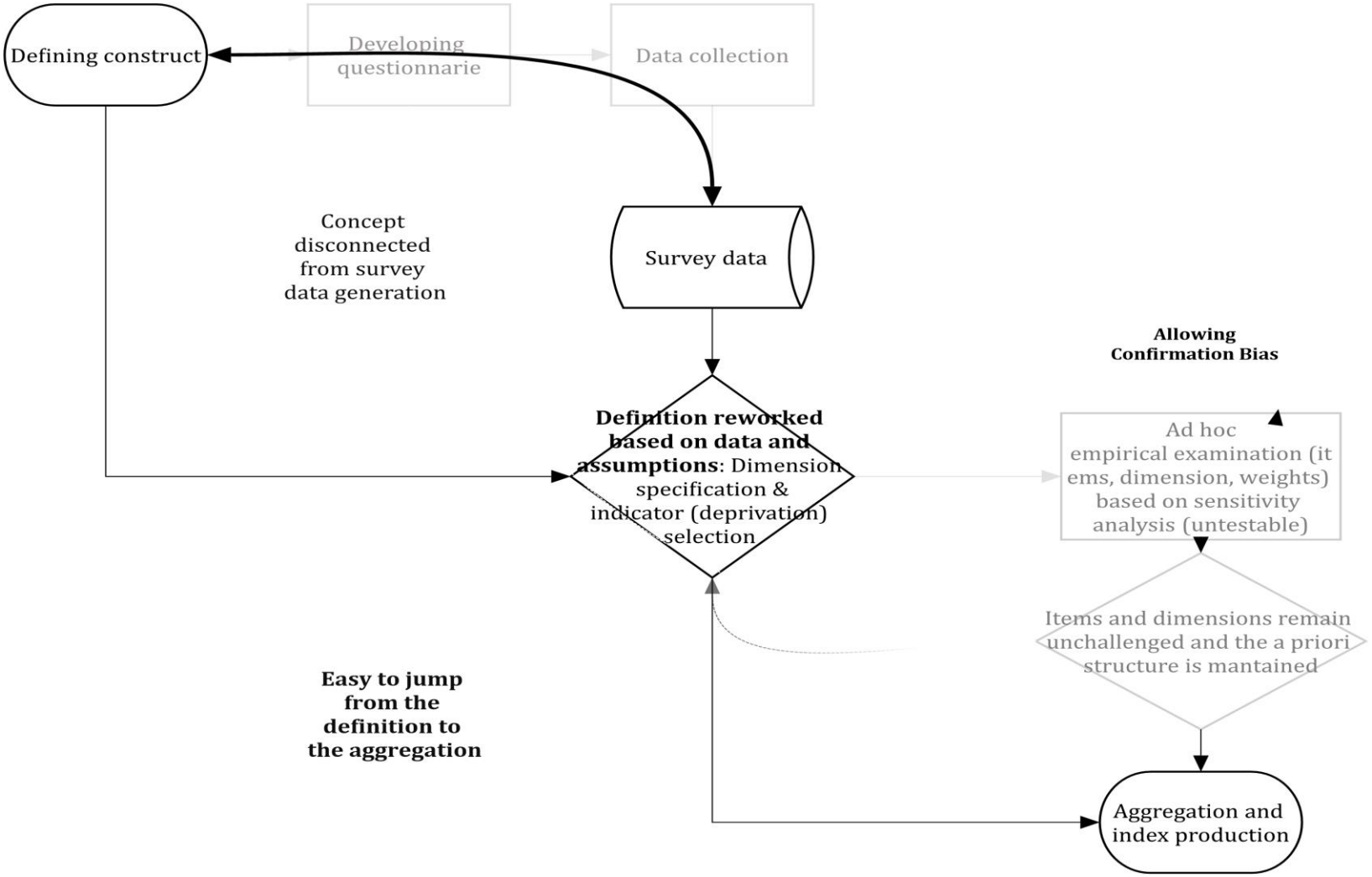
Is the ideal framework implemented in practice?

What do you think?

- Not very often. The only example that fully covers all the steps is the relative deprivation theory and measure
- Relative deprivation theory > Consensual module > Empirical scrutiny
- Capability > ? > Available data > Axioms
- What do we se in practice?

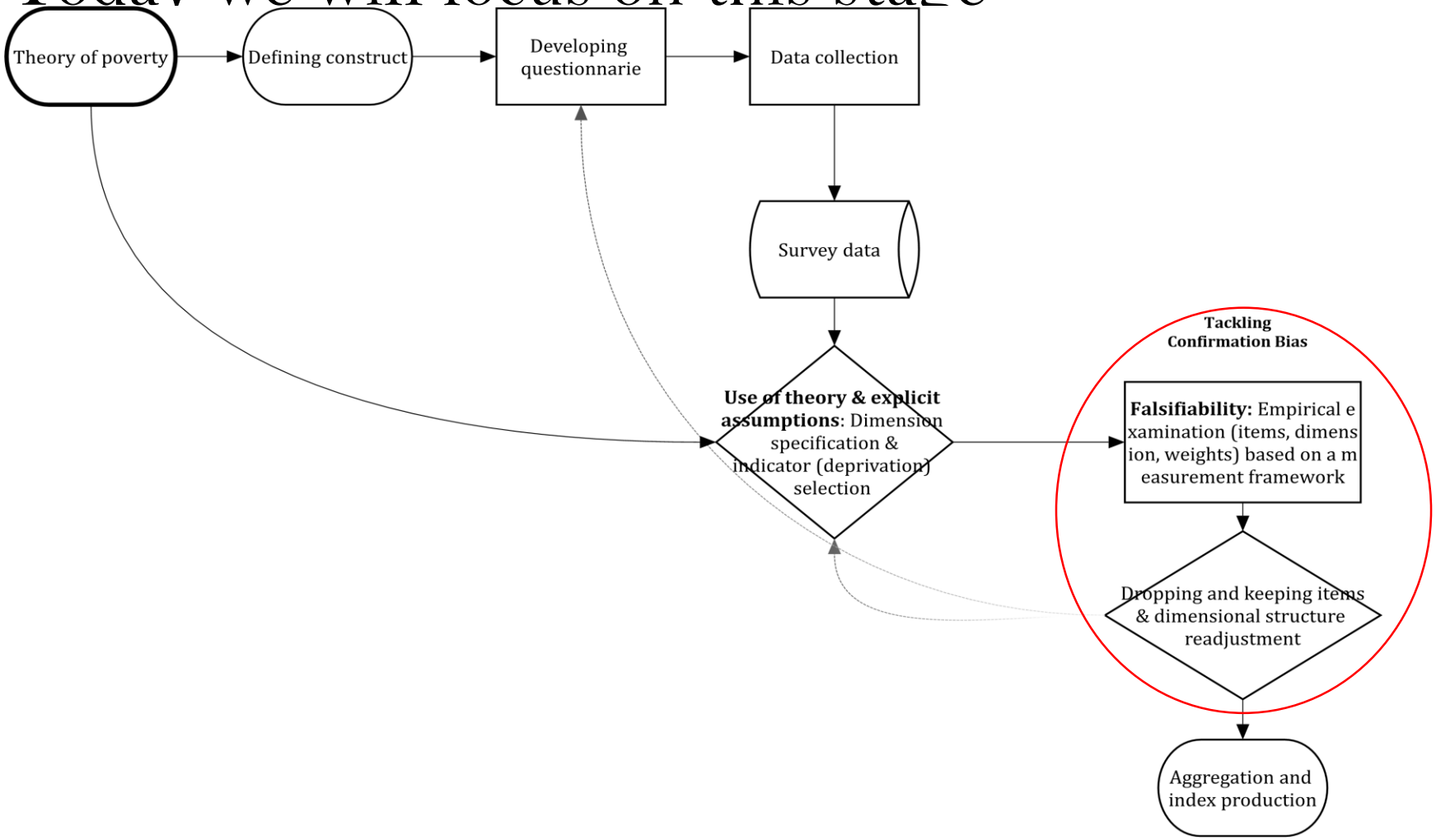
Real workflow in poverty measurement

The next figure illustrates the strategy often followed by researchers to produce a poverty index.



The workflow researchers often implement in multidimensional poverty measurement.

Today we will focus on this stage



Ideal workflow in multidimensional poverty measurement

Measurement is befouled by assumptions and error

- The perfect measure is unknown and researchers put forward a theory to make poverty trackable via data
- Given a definition of poverty, there are different dimensions, indicators and parameters to produce a working model to approximate poverty.
- Because we approximate poverty by raising a series of assumptions, we need a measurement theory that tells us **from an empirical perspective** whether our guessing exercise leads to sensible results

Scientific measurement in social sciences

- There is no such thing like **error-free measurement**. Even less so when we are using concepts that we think bear some relation with reality
- Researchers will hardly disagree with this (although there are some!) and yet measurement is not taken seriously enough
- As Loken and Gelman (2017) point out: The replication crisis in social sciences is to a large extent due to poor measurement practices
- Random noise leads to random conclusions

University of BRISTOL

Universidad Nacional Autónoma de México UNAM

Scientific measurement in social sciences

- In social sciences we work with concepts, i.e. abstractions or things we think exist but are not directly observable
- So how do you measure a concept?
- One framework to do so is measurement theory or, more precisely, **latent variable theory/modelling**
- This measurement theory has more than 100 years of continuous development and yet its implementation in poverty research is fairly recent
- The most well-known implementation so far is the EU material deprivation index

Measurement theory and principles

Poverty measurement requires some governing principles that effectively put the workflow in poverty measurement in terms of a cogent falsifiable framework.

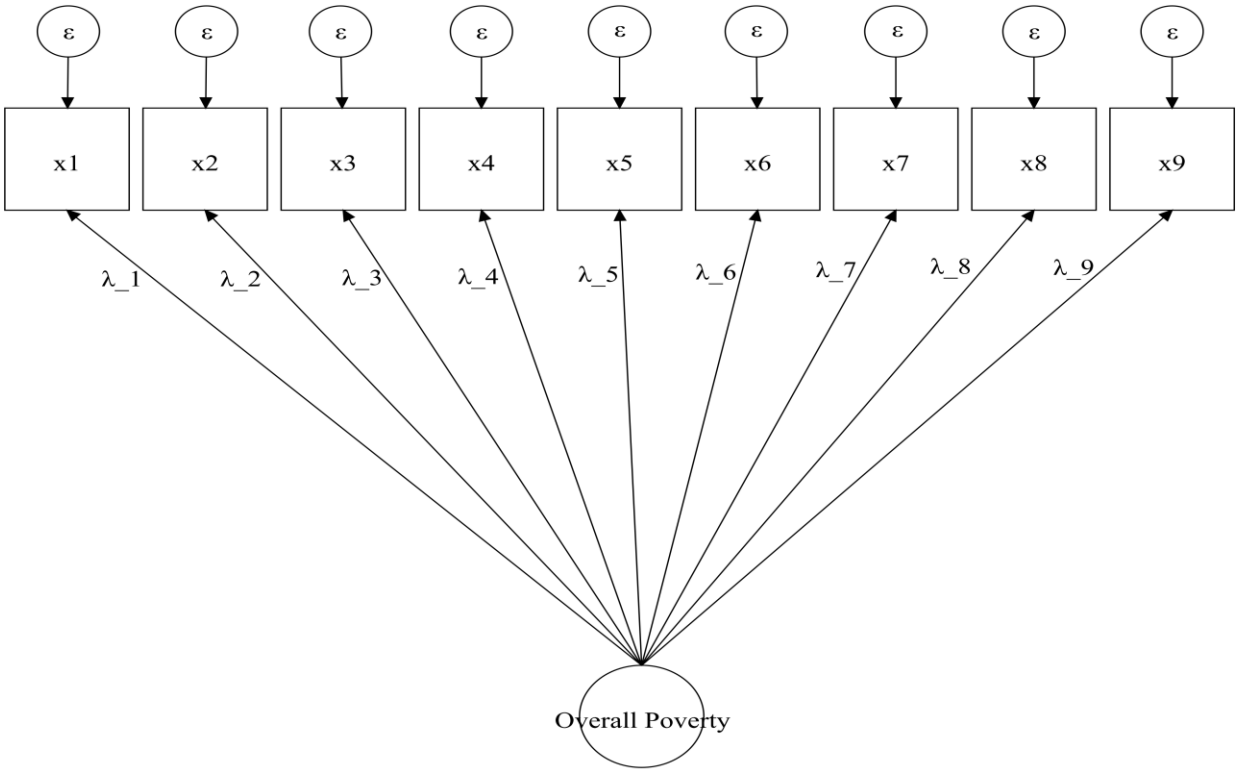
- Is the subset of dimensions j from $\mathcal{J}$ an adequate characterization of poverty?
- Is the subset of indicators given a cut off $(X; z)$ from $\mathcal{X}$ an *adequate/good* characterization of the dimension j and poverty?
- Does the weighting scheme lead to the same ranking of the population?
- Does the selected poverty line leads to a meaningful split of the population?

University of BRISTOL

Universidad Nacional Autónoma de México
UNAM

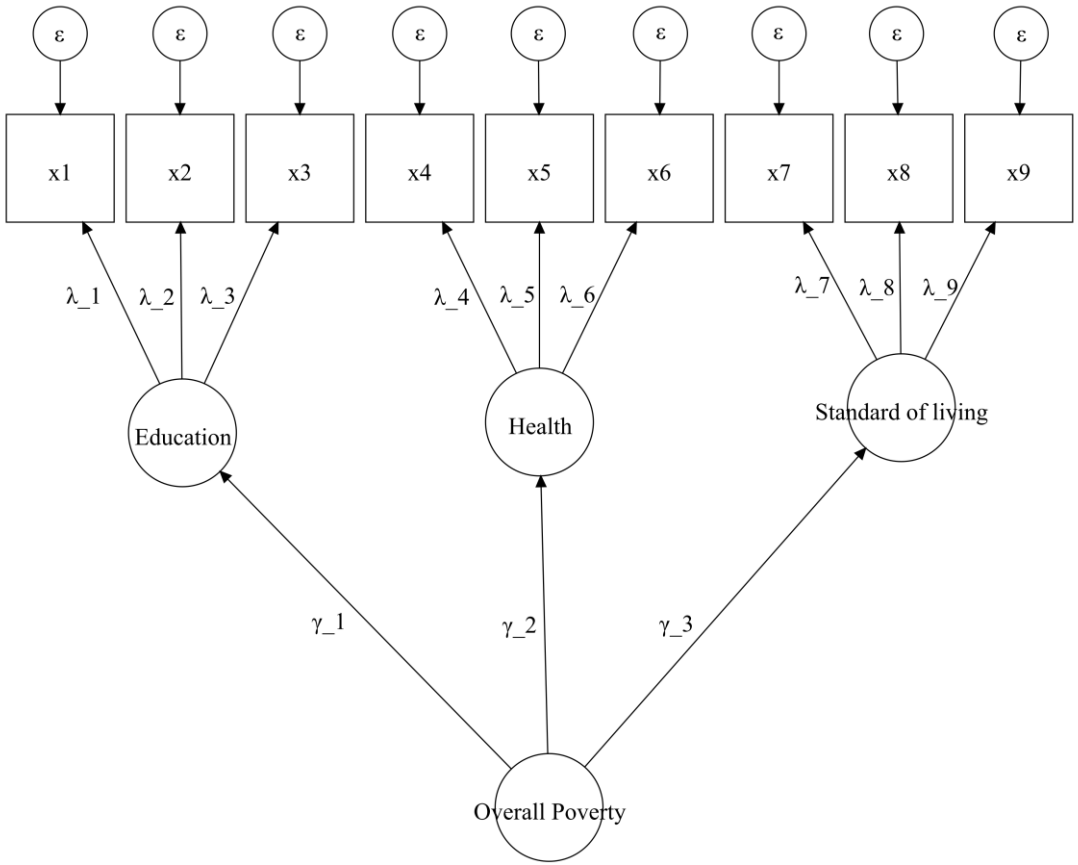
In poverty research we have blueprints (how our measure looks like)

A unidimensional model



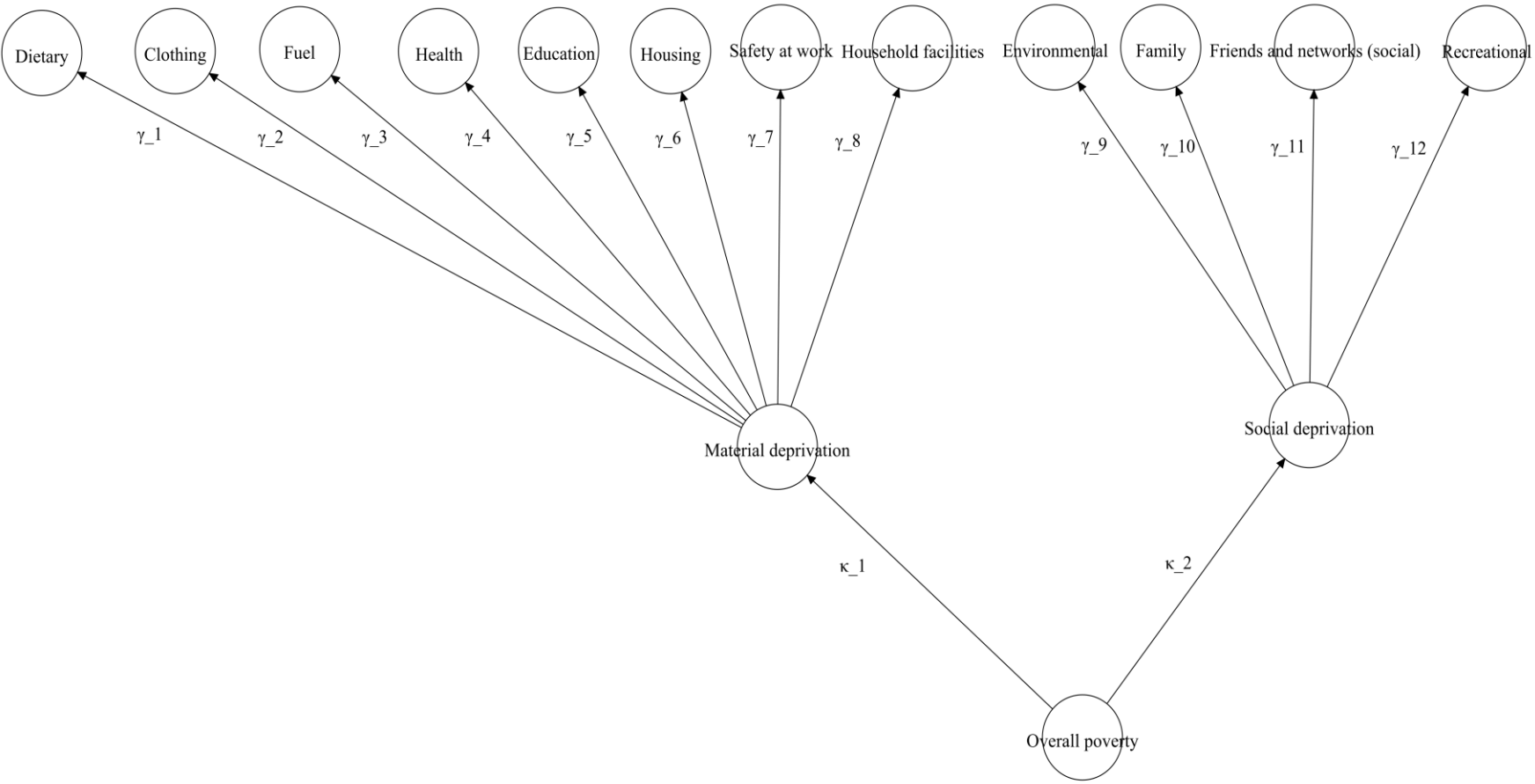
This is a visual representation of a null unidimensional model.

The UNDP-Oxford idea



This is a visual representation of Alkire and Santos (2010)'s model. Second-order factor

Townsend's idea



This is a visual representation of Townsend (1979)'s model. Third-order factor.

What are the key questions then?

- We want to know which model is an adequate representation of poverty, i.e. are valid and reliable
- What dimensions seem to hold given the data?
- Which indicators provide a good account of the dimension in question?

Spearman's theory of latent variables

Spearman (1904) put forward two capital ideas:

- We can't observe a concept directly but we can measure its manifestations
- Remember that according to Townsend (1979):

Poverty is the lack of command of resources over time and material deprivation is its consequence

Deprivation/achievements are the observed outcomes of poverty

Spearman's theory of latent variables

- The second main idea of Spearman (1904) is:
  - Two things are correlated because probably they are caused by the same thing
- So a latent phenomenon can be captured via its correlated manifests or outcomes.

University of BRISTOL

Universidad Nacional Autónoma de México UNAM

How do we define **good** measurement?

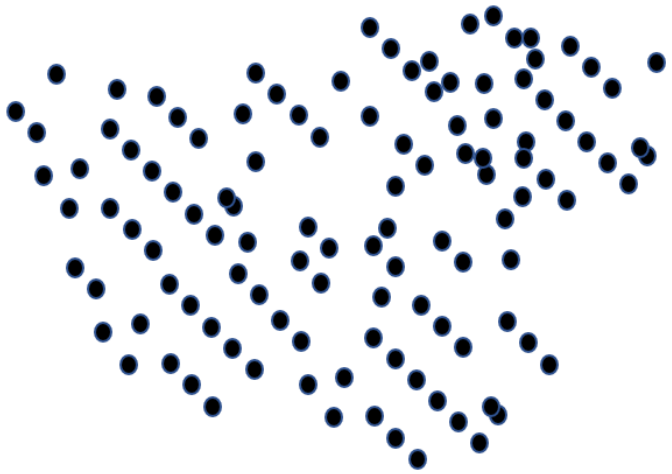
Measurement theory since the seminal work of Spearman (1904) has continuously developed a cogent framework that aims to produce measures that do the following:

University of BRISTOL

Universidad Nacional Autónoma de México
UNAM

University of BRISTOL

Universidad Nacional Autónoma de México
UNAM



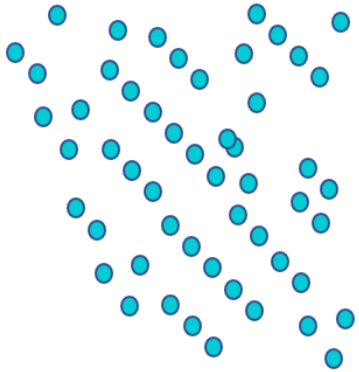
Population

University of BRISTOL
Universidad Nacional Autónoma de México UNAM

Population classification

University of BRISTOL
Universidad Nacional Autónoma de México
UNAM

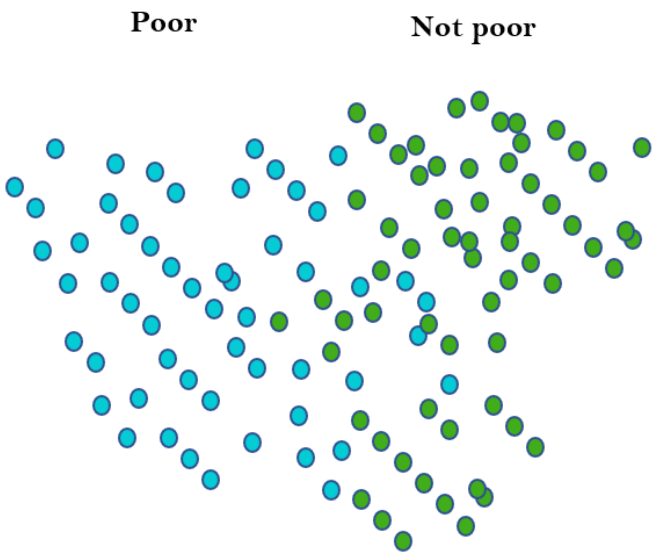
Poor



Not poor

Ideal classification

University of BRISTOL
Universidad Nacional Autónoma de México UNAM



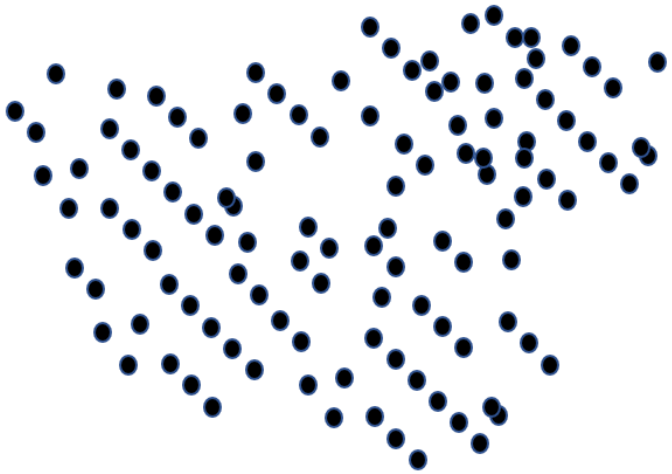
Real classification

University of BRISTOL

Universidad Nacional Autónoma de México
UNAM

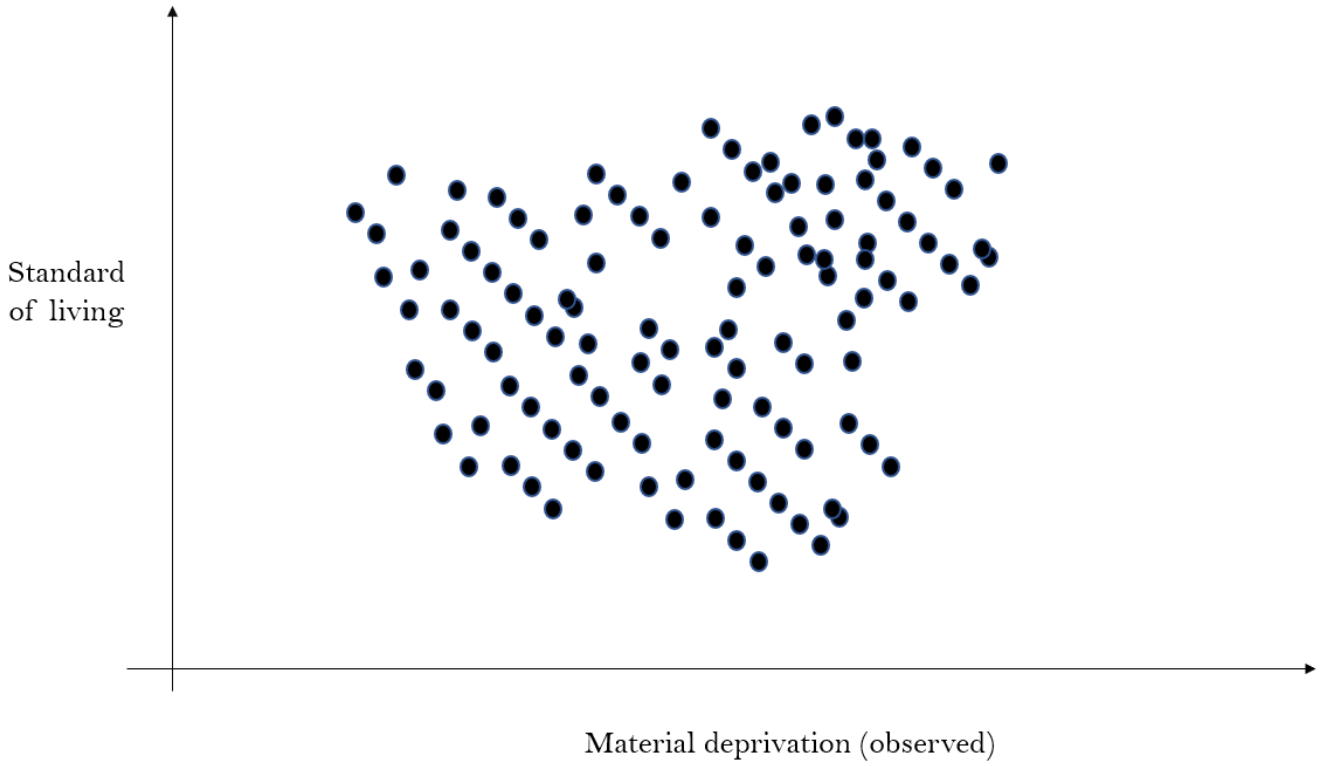
University of BRISTOL

Universidad Nacional Autónoma de México
UNAM

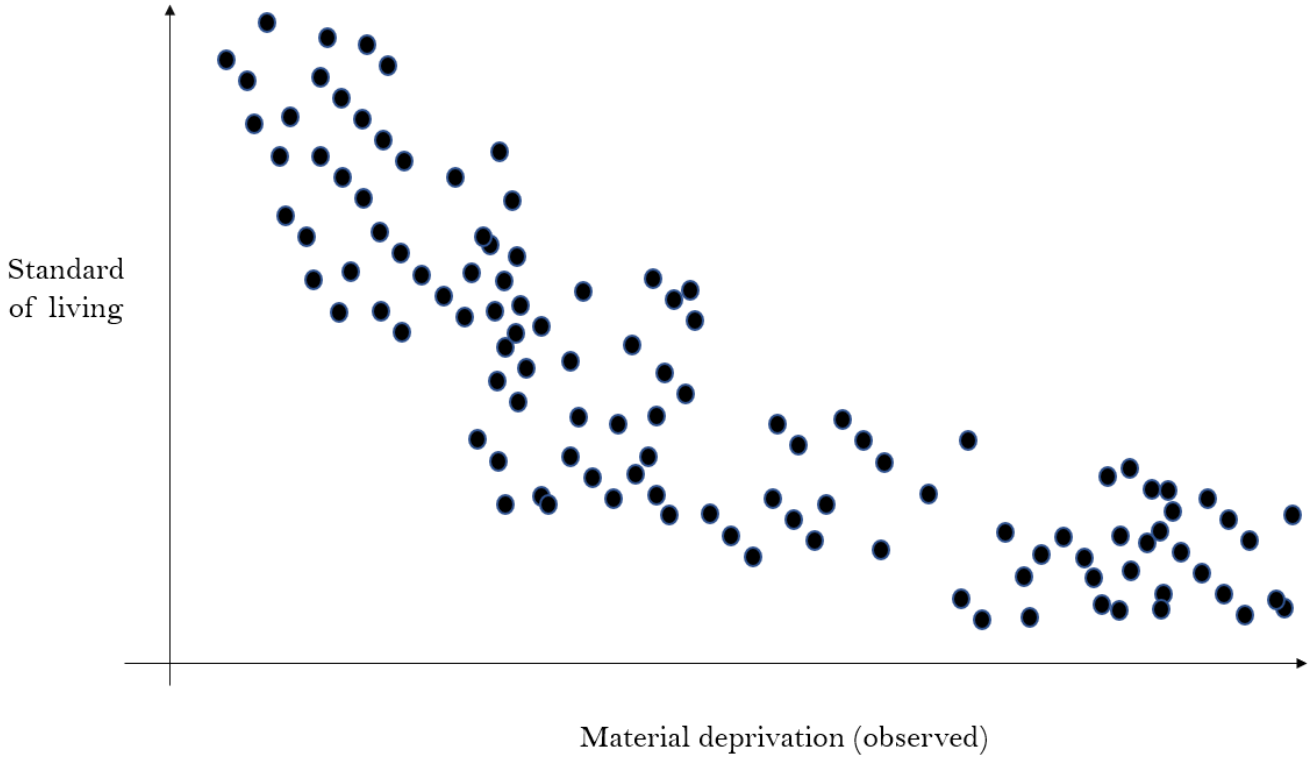


Population

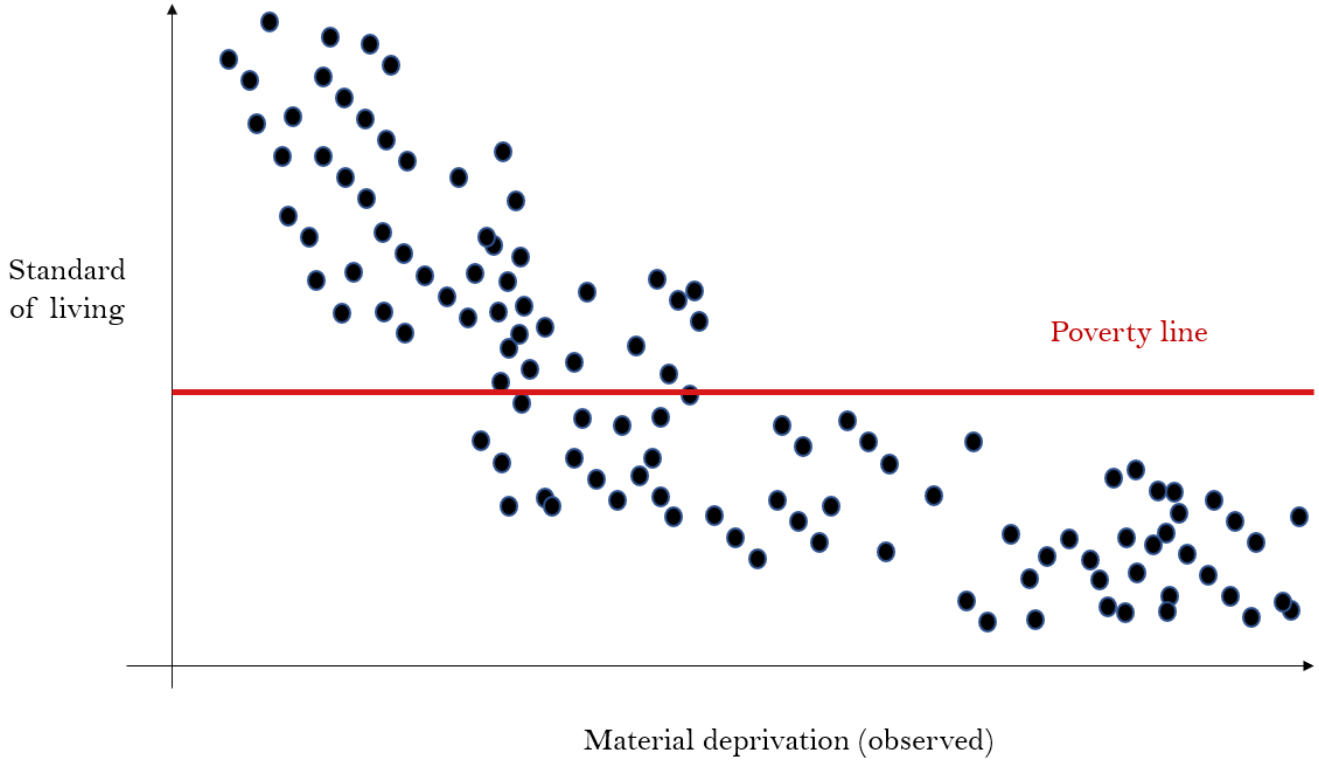
University of BRISTOL
Universidad Nacional Autónoma de México UNAM



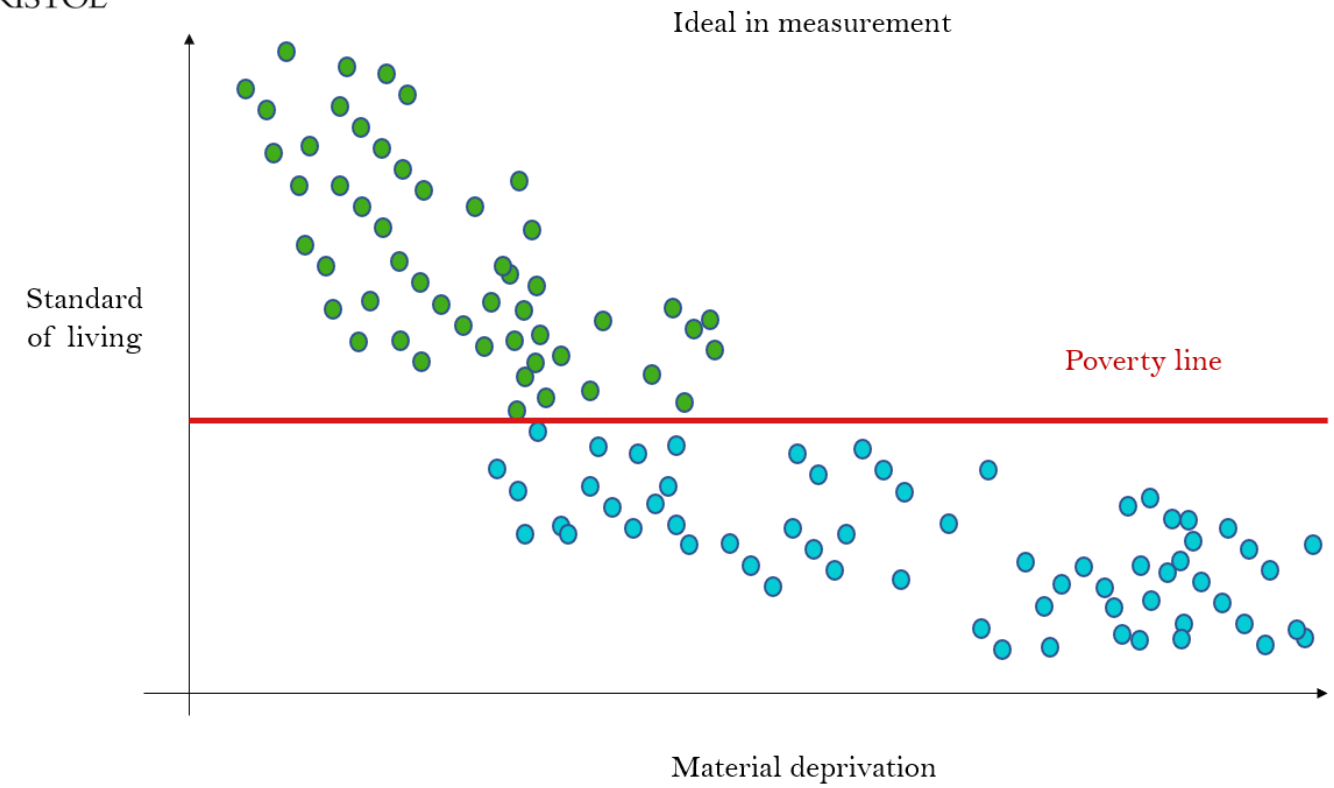
University of BRISTOL
Universidad Nacional Autónoma de México
UNAM



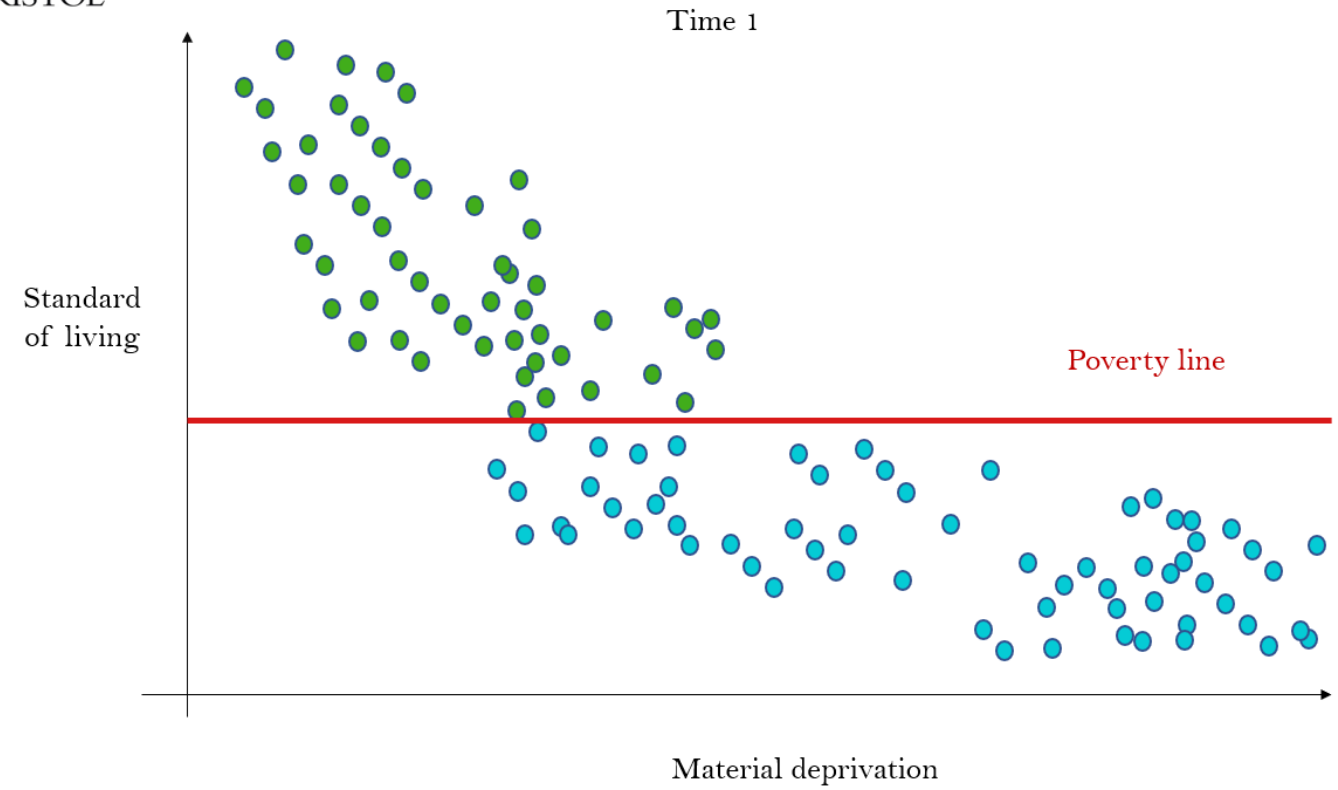
University of BRISTOL
Universidad Nacional Autónoma de México UNAM



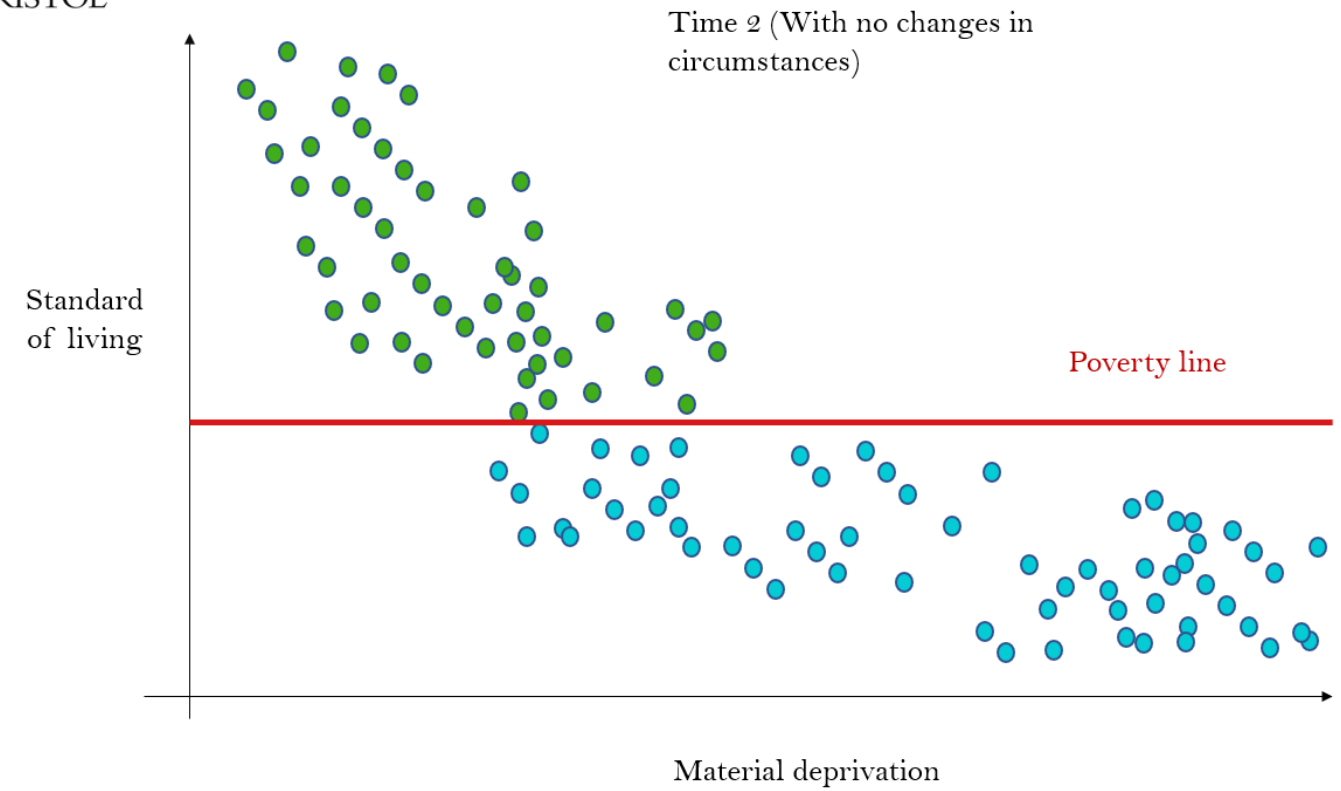
University of BRISTOL
Universidad Nacional Autónoma de México
UNAM



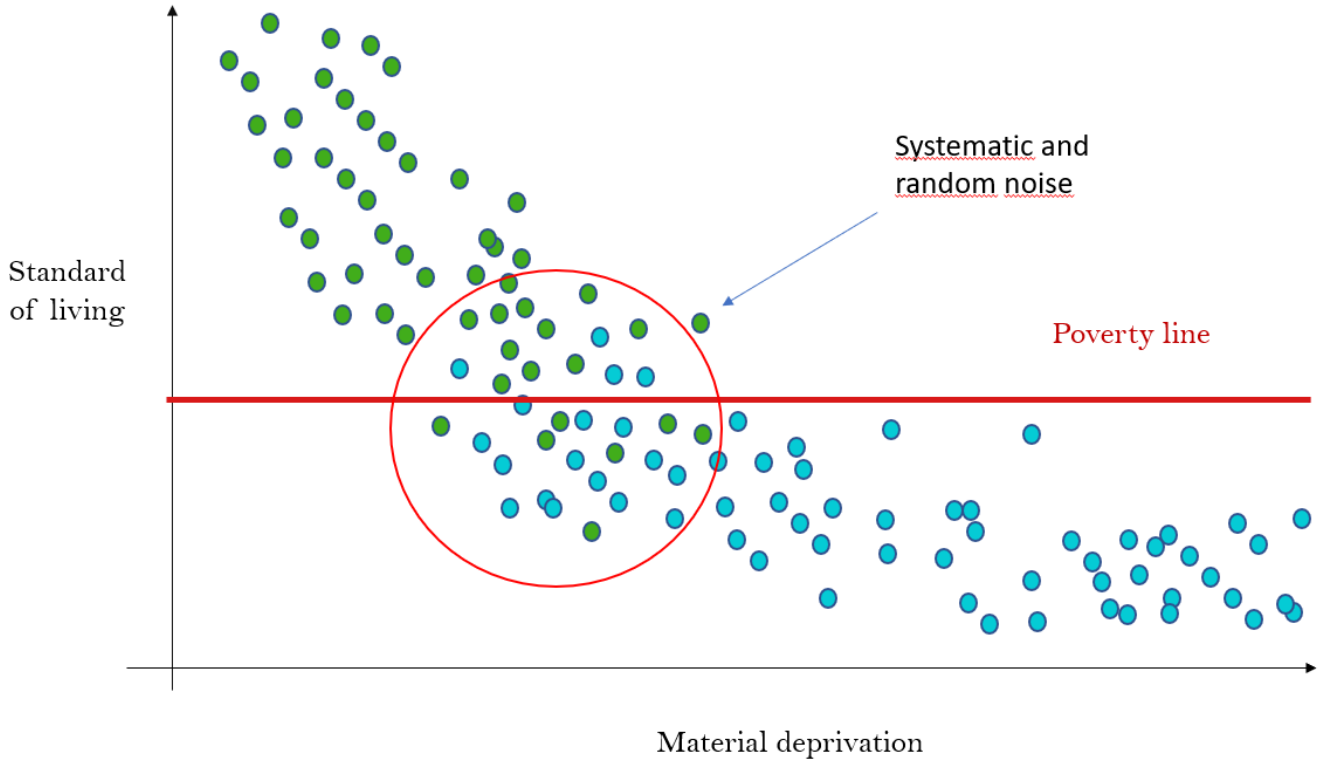
Reliability leads to consistency across measurement as follows



University of BRISTOL
Universidad Nacional Autónoma de México
UNAM

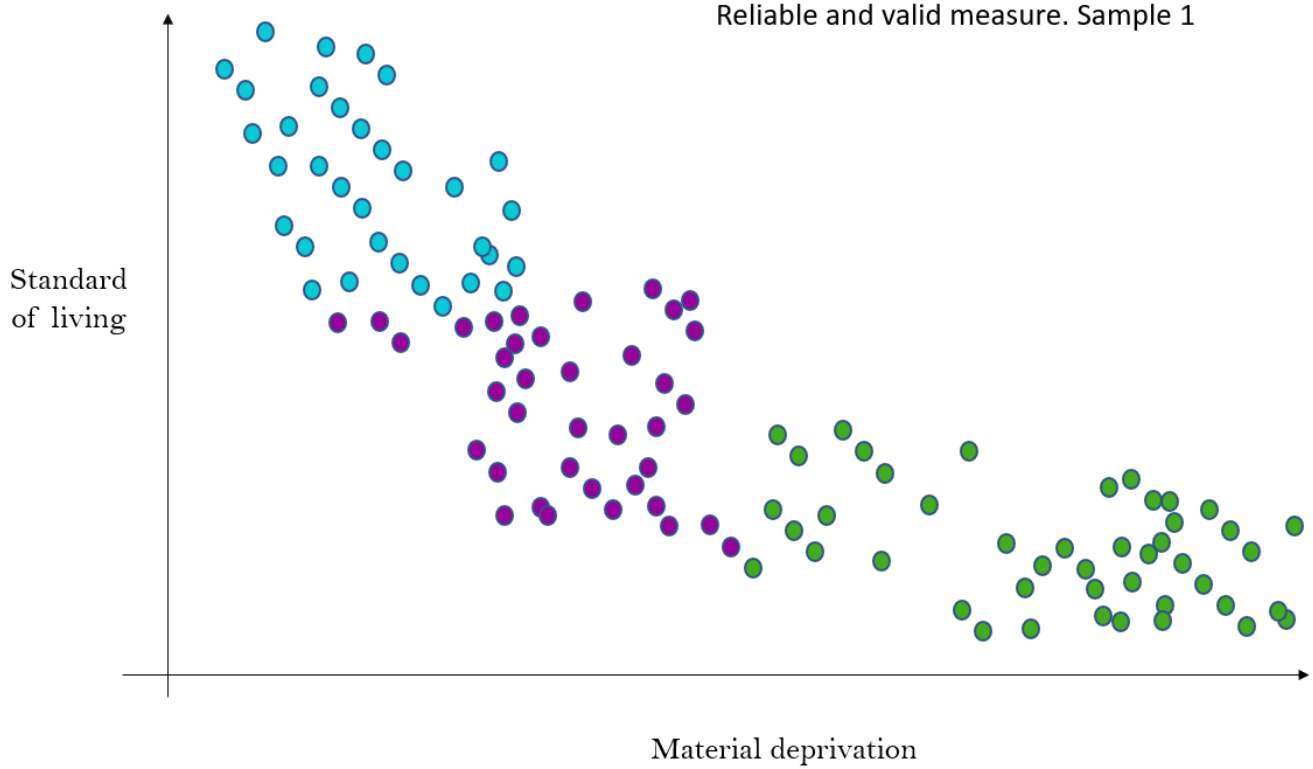


But we don't have perfect measures, we always have noise!



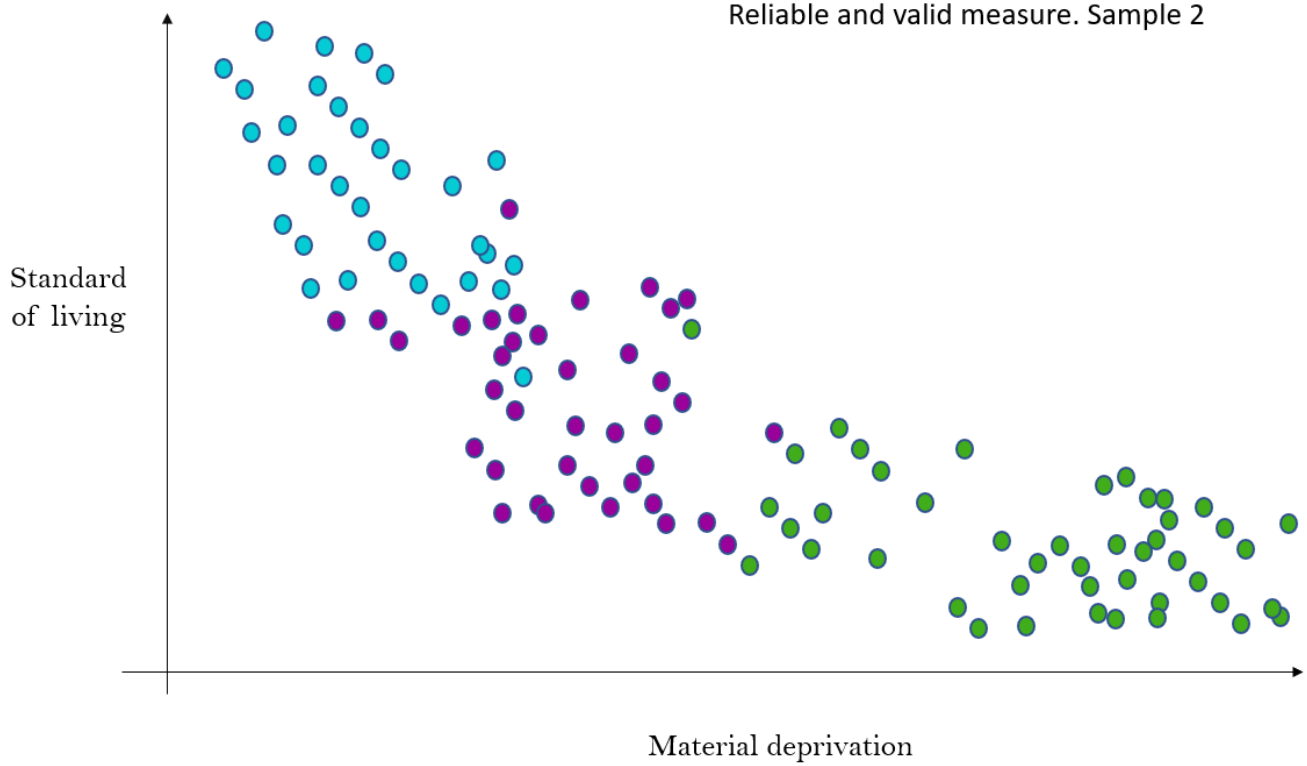
University of BRISTOL
Universidad Nacional Autónoma de México UNAM

Reliable and valid measure. Sample 1

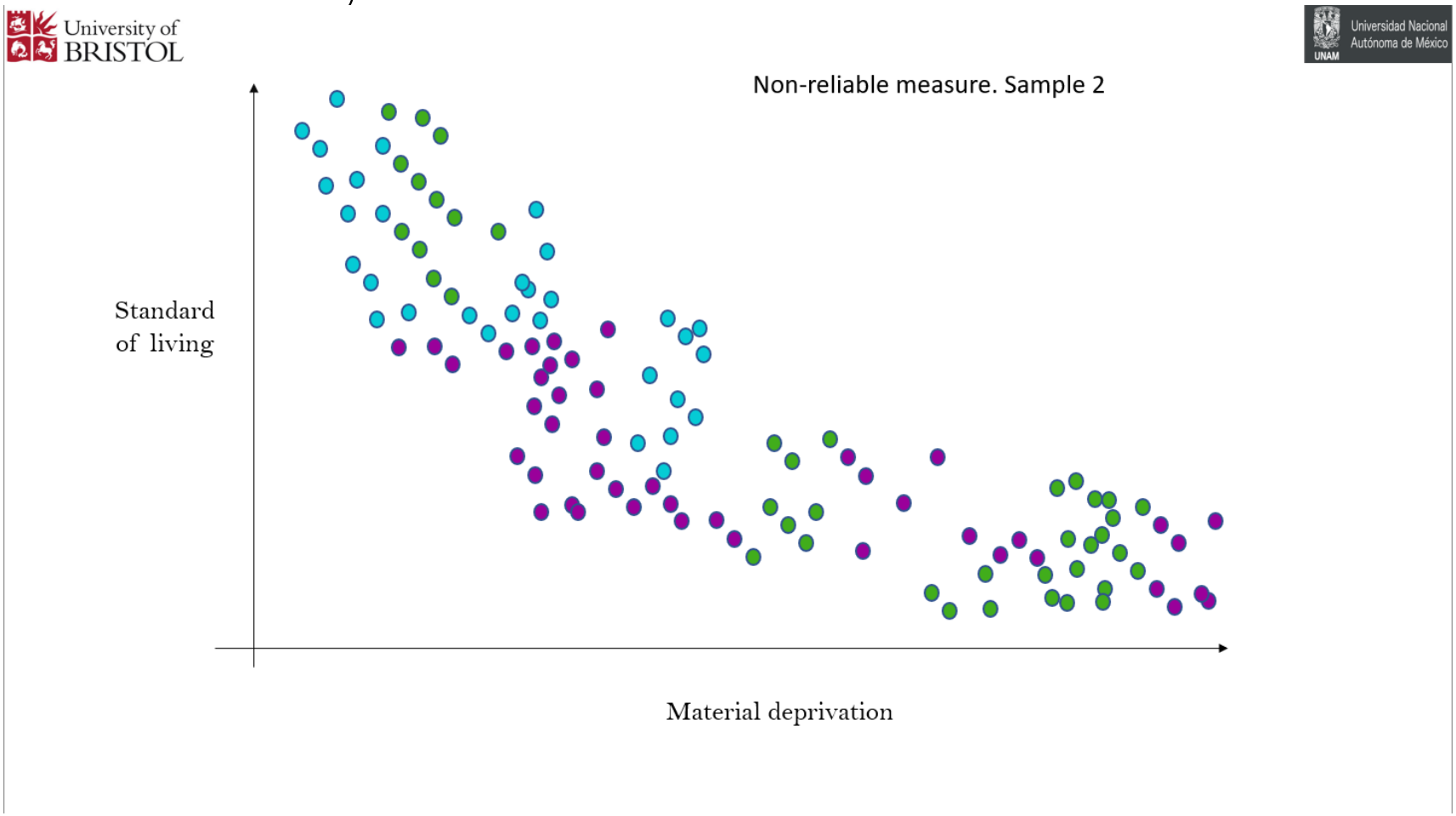


University of BRISTOL
Universidad Nacional Autónoma de México UNAM

Reliable and valid measure. Sample 2



This is what you want to avoid in measurement, which is a consequence of violating reliability and validity



To sump up measurement theory thus aims to produce measures that:

1. Consistently offers the same ranking of a population
2. A ranking represents an ordering of the population with respect the phenomenon we aspire to measure. The two aims gave birth to the concepts of *reliability* and *validity*.

These two have important implications in terms of weighting, comparability and identification of (unobservable) population groups.

Intuition to the concept of reliability

- There is nothing worse in measurement that a scale that causes disbelief in that the debate concentrates upon how bad a measure is and not upon how good or bad a policy is.
- `Trust' is built upon consistent and meaningful estimates. For example, imagine a case in which we could conduct two surveys to the same population. Ideally, we expect to find that the response patterns remain unchanged from t_1 to t_2.
- A noisy index, in contrast, would lead to unstable responses and it is impossible to distinguish a signal (the thing we are interested in) from noise (unnecessary and confusing variability).
- However, consistency is not simply having the same response patterns *ceteris paribus* across two samples. The main implication for measurement of consistency is that we will get systematic population orderings, i.e. people ranked with very low (latent) living standards should remain in the same position across different measurements

Reliability and bad indicators

- Imagine a case in which one of the deprivation indicators is not a good measure of poverty, like having a folding bicycle.
- This variable will have a low correlation with the rest of the deprivation indicators. Spearman (1904)'s theory of reliability (or attenuation) tells us to be suspicious about such kind of behaviour.
- The problem is that low correlation (or even worse negative correlation) could mean that the indicator in question in not a consequence by poverty ("Lack of command of resources over time").
- The effect of bad indicators is that we will end up with two different population rankings depending on whether we include folding bicycle in our index.
- How different? It will depend upon how poorly correlated the indicator in question is with the rest and how good the rest of indicators is as a whole -i.e. how reliable these are-. Therefore, even with a very similar response pattern, our scale will be rather unstable to be trusted.

Realiability and good indicators

- Now imagine a different scenario where we have only good outcome measures of poverty and, for some reason, a good variable like lacking drinking piped water inside the house is dropped from the index (assuming this is a developing country where this measure works!).
- If we drop this indicator from our analysis, we would lose valuable information.
- Lack of good quality data or bad theories will increase the risk of missing some good indicators. Therefore, we would like that our scale is protected against the perils of real-data analysis. In other words, we would like to have a measure that is not too sensible to information losses.
- High reliability is a property that, for instance, protects an index against certain information losses, i.e. the higher the reliability, the lower the effect of missing variables. Yet, missing indicators could be damaging for policy reasons, of course.

Formal introduction to reliability

- Reliability is a key concept in measurement theory and can be simple defined as the homogeneity of an index (Revelle and Zinbarg 2009).
- A homogeneous index is a scale whose indicators are manifestations of the same trait, which is why from Spearman (1904)'s theory we expect them to be correlated as they are caused by the same phenomenon
- In the literature, several authors refer to reliability as internal consistency of an index because this a consequence of homogeneity.
- How this term of homogeneity relate to our intuition about reliability? In the example above having an indicator that is not a good measure of poverty means that the index is heterogeneous -i.e. there is more than more phenomenon causing deprivation- and therefore leads to inconsistent population orderings.
- Thus at the core of the principle of reliability lies the idea of having a series of items that would have a predictable behaviour when aggregated, i.e. if an index is reliable we should expect to have very similar population rankings across samples or small variations of the same reliable index with more or less indicators.

Estimates of reliability

- There are different ways to estimate the reliability of a scale, each one with its advantages and disadvantages.
- The most widely used estimator/index of reliability is α or λ_3 (do not mistake with factor loadings) (Cronbach 1951; Guttman 1945).
- This estimate comes from Clasical Test Theory (true + error)

- Draws upon Spearman (1904) approach to estimate the variance based on parallel tests and, more importantly, the idea that by correcting the value of a correlation by *attenuation* would lead to the best estimate of reliability. In other words: Are the observed correlations the true correlations? How this relationship is attenuated by noise?

Do not **only** use alpha!

- Cronbach's α is, nonetheless, not a good estimate of reliability (Zinbarg et al. 2005; Revelle and Zinbarg 2009).
- It only works fine under very restrictive assumptions. First, the association between each indicator and the latent variable is equal.
- For example, for a measure based on three outcome variables it would mean that: $\lambda_1 = \lambda_2 = \lambda_3$. Second, it assumes that the variances across tests are equal.
- These two assumptions are, however, necessary because otherwise it is not possible to compute σ_θ^2, σ_x^2 and the covariances.
- These two assumptions are unlikely to hold in practice. Another problem with α is that increasing the number of items and the average inter-item correlation will increase the reliability estimate.

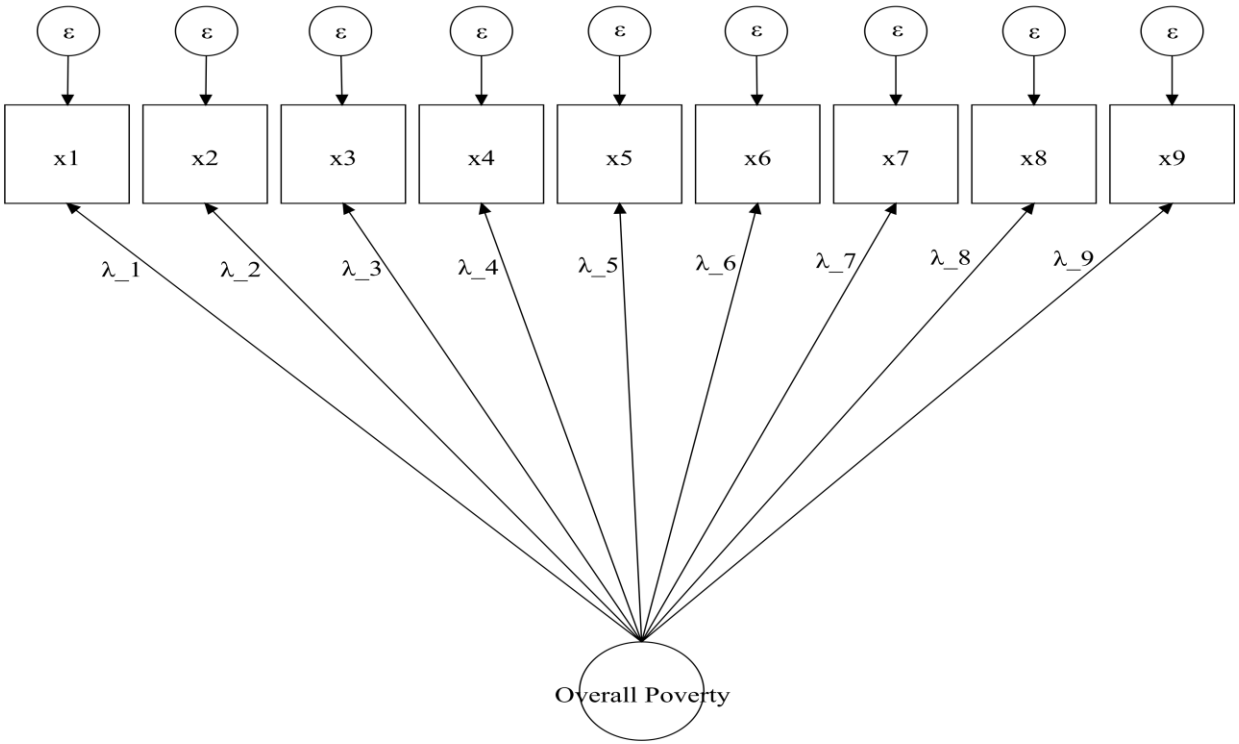
Use Omega and Omega_h!

- Moving beyond CTT, McDonald (1999) put forward two alternate measures of reliability: ω and ω_h.
- The statistics can be better framed within latent variable modelling and thus are estimated using factor analysis - preferably confirmatory factor analysis (Brown, 2006).
- Both use the idea that the variance of the outcome measure accounted by for the factor for each indicator and then aggregates this to produce indices for the scale as a whole.

Omega

- The first statistic (ω) is also know as the measure that maximizes the estimation of reliability, i.e. the lowest upper bound (Zinbarg et al. 2005).
- ω is a proportion of the variance of the outcome measures that is explained by the factor.
- In other words, if we have several very good deprivation indicators that are caused by the lack of command of resources over time, we expect the error to be low and the loadings of each indicator to be very high.
- Consequently, ω will be very high, i.e. close to 1, which is its maximum value.

To put differently



This is a visual representation of a null unidimensional model.

Omega h

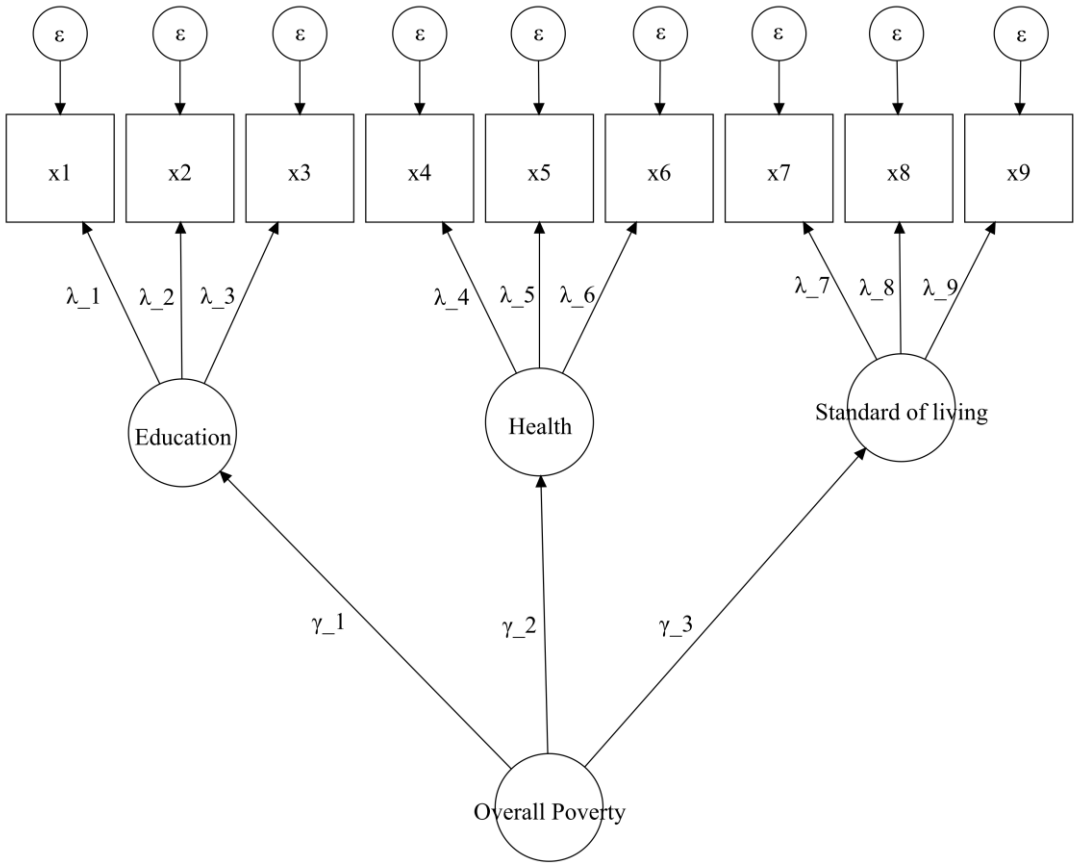
- The statistic ω focuses on the unidimensional case in that it main concern is working out the percentage of the variance explained by the factor.
- However, in multidimensional poverty measurement we have a higher-order factor and several nested factors (dimensions of poverty). Thus, there we have at least to sets of relationship in the multidimensional case:

(1)The relationship between each outcome measure and the overall factor (poverty) and,

(2)the relationship between each dimension and its outcome measures. In fact, this is more complex that it seems, as the first relationship can be also put in terms of the relationship between the higher-order factor and each dimension.

The answer to the problem of nested dimensions for the computation of reliability is ω_h. It is a hierarchical version of ω in that it aims to estimate the two sets of relationships stated above: the variance of the indicators explained by both the higher order factor and the subdimensions.

To put differently… you see how big the error is



This is a visual representation of Alkire and Santos (2010)'s model. Second-order factor

Are there any thresholds to evaluate reliability?

- One of the consequences of reliability is that it leads to an accurate ranking or ordering of the population in question, i.e. from the lowest standard of living to the highest.
- Nájera (2018) run a Monte Carlo study to assess the relationship between reliability and population classification. Hence, this study poses the question about the level of reliability that guarantees a low amount of error.
- The result was that there is a clear relationship between reliability and population classification.
- For unidimensional measures: $\omega > .8$ leads to an error of $< 5\%$ and entropy $> .8$
- For weak-dimensionality: $\omega > .85$ and $\omega_h > .65$ leads to an error $< 5\%$ and entropy $> .8$
- For strong-dimensionality: $\omega > .85$ and $\omega_h > .70$ leads to an error $< 5\%$ and entropy $> .8$

Item-level reliability

- Classical test theory was concerned with overall reliability.
- Item response theory (IRT) moved from the idea of a true score and look at the relationship of the indicators with an underlying trait (e.g. intelligence, depression, poverty) (Harris 1989).
- IRT is a theory about the type of relationship that an indicator has with a latent variable. The simplest IRT specification proposes that a measure is unidimensional (i.e. the variance of the indicators is accounted by for one trait) and that each item relates to different degrees of difficulty or severity of the construct. This is called a one-parameter IRT model.
- A more general IRT model also proposes that some indicators are better than others to differentiate the population. That is, that some deprivation indicators are associated with a higher likelihood of belonging to the poor group. This more general aspect is added via a second parameter called discrimination and leads to a two-parameter IRT model.

Item-level reliability. IRT and CFA

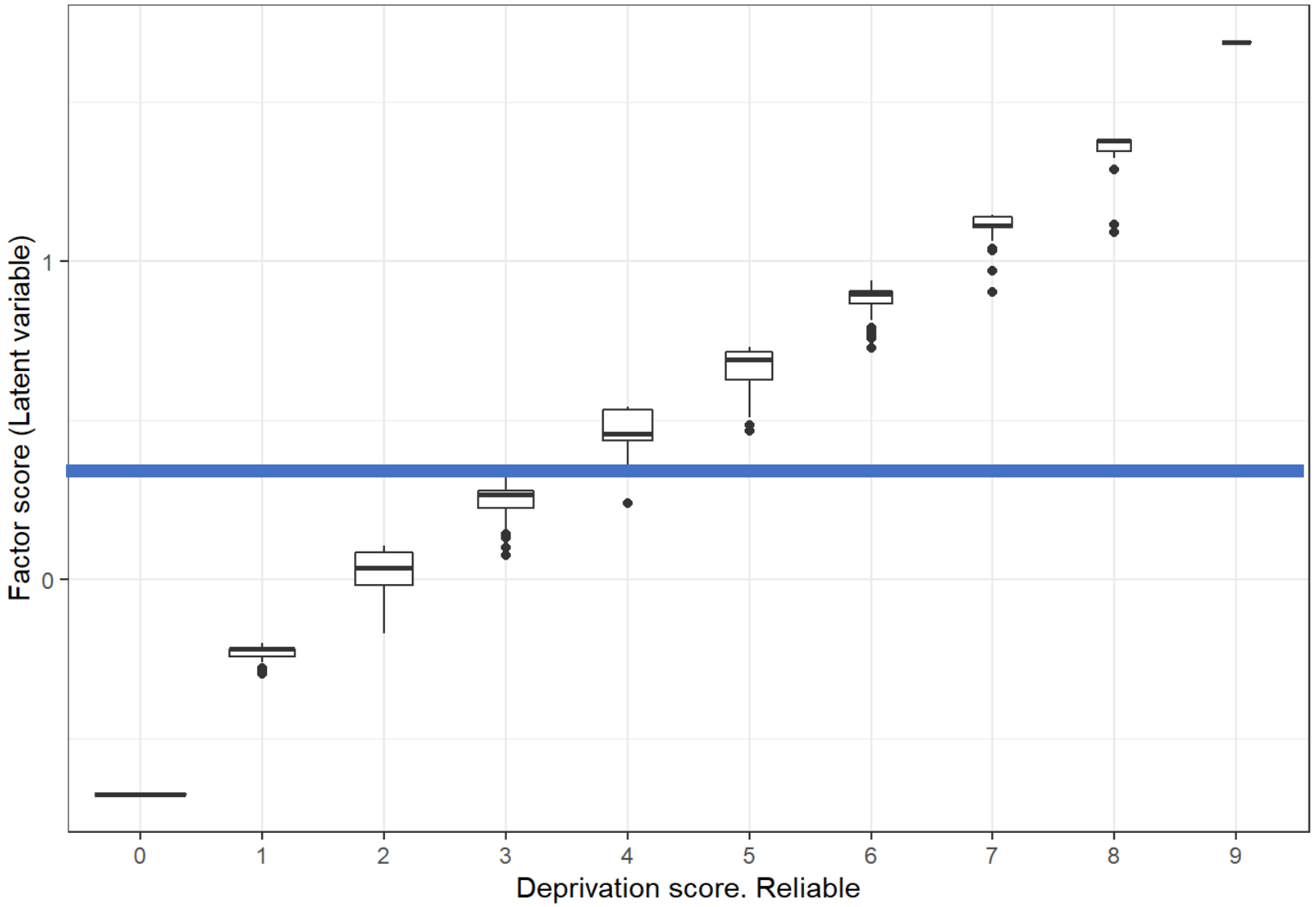
- Translated to poverty measurement, IRT modelling states that the probability of choosing a someone that is deprived in the indicator i is given by the discrimination (a) and the severity(b) of the item.
- Muthén (2013) shows that (b) is just a threshold and (a) the factor loadings (λ_i).
- Therefore, the stronger the loadings, the higher its discrimination power, where ψ is the variance of the latent variable.
- The original IRT models work under the assumption of unidimensional scales, i.e. one factor with several manifest variables that exclusively belonged to such factor.
- Gibbons et al. (2007) have shown that the presence of a higher-order factor produces little bias in the estimates when having more dimensions. In theory, all multidimensional poverty models make such an assumption. In any case, the concepts remain the same and a multidimensional IRT model can be simply connected with multidimensional confirmatory factor model.

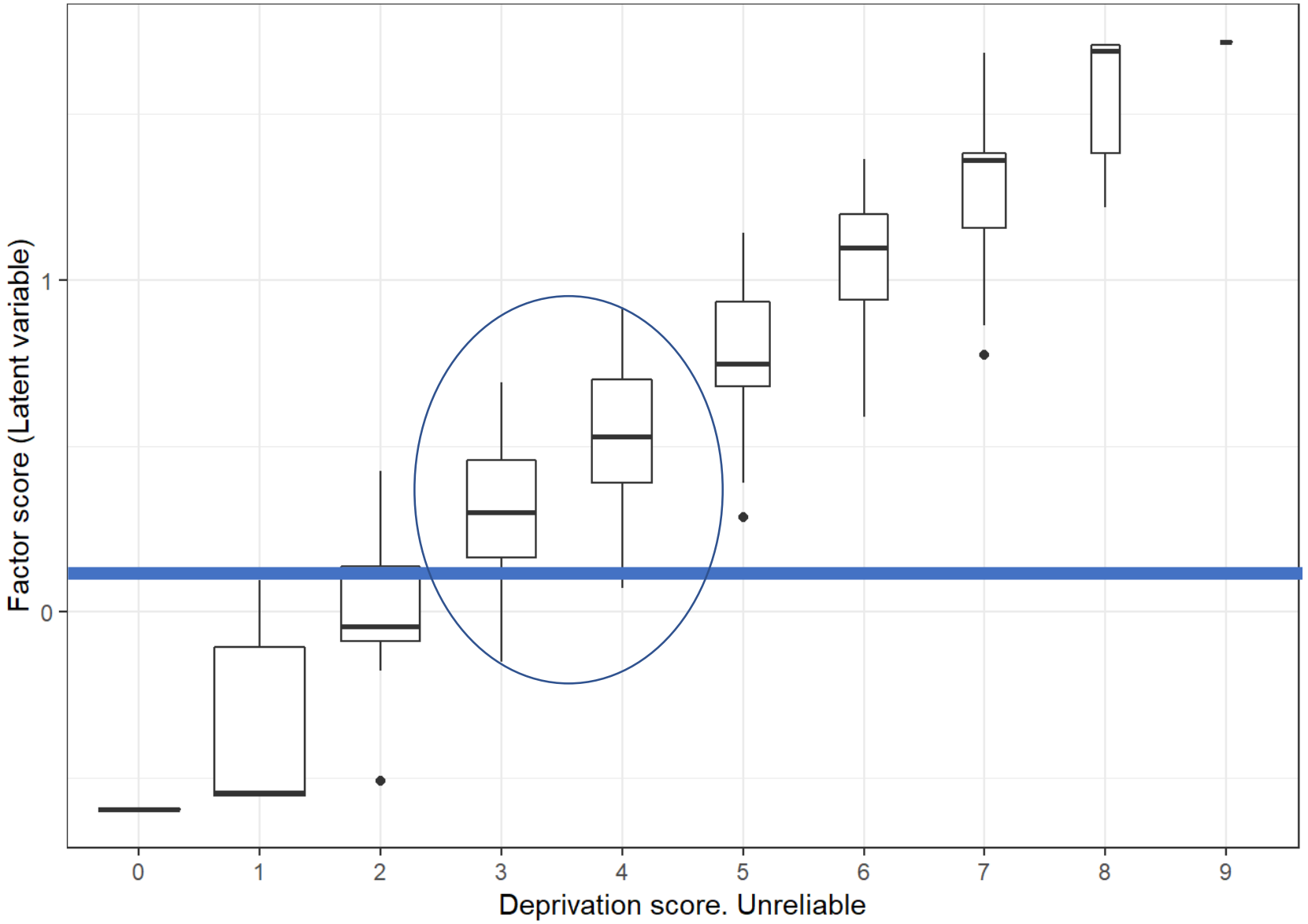
Thresholds for item-reliability

- Statistics such as β, α, ω provide an summary of the overall reliability.
- The computation of ω heavily relies on the factor loadings. The lower the factor loadings the higher the error and the lower the overall reliability.
- Similarly, low λ_i can be translated as low item-level reliability values.
- The question is thus how low mean unreliable. Guio et al. (2016) use the rule of $< .4$ standardised loadings (or $<.8$ unstandarised/Discrimination) as a measure of item-unreliability. Nájera (2018) shows that indeed those values are more likely to result in overall unreliability and high population classification error.

Reliability, population orderings and identification error

- One of the most contested issues in poverty measurement revolves around weighting (Decancq and Lugo 2013).
- Measurement theory proposes that reliability lead to a self-weighting measure in that it guarantees good population classification (Streiner, Norman, and Cairney 2015).
- Discrimination parameters have a crucial role upon population classification and item weighting. The square of the factor loadings equals the amount of variance in the indicator explained by the common factor (i.e. communality).
- Because the factor loadings capture the relationship of each indicator with the latent variable, they can be seen as the optimal weights of the model given the data.
- Therefore, a test of equality of loadings within dimensional can be used to assess whether using such kind of weighting is reasonable or not.
- Nájera (2018) shows that very high reliability leads to a self-weighting index in that the population ranking is less sensible to the items used in a scale.
- Therefore, discussing the use of differential weights versus non-differential weights misses the point. The critical point is that differential weights, in that they are unknown, will always introduce more noise to the classification of the population. Whereas reliability is a necessary condition for good population orderings, weighting it is not so.


Factor score (Latent variable)
1
0
0
1
2
3
4
5
6
7
8
9
Deprivation score. Reliable


Factor score (Latent variable)
1
0
0
1
2
3
4
5
6
7
8
9
Deprivation score. Unreliable

Moving from reliability to validity

- Reliability is homogeneity in measurement and it means the capacity of a measure to reproduce the ranking of a population under changing conditions.
- Reliability will tell us whether the set of indicators will be useful to order individual's according to their latent scores which we **presume** reflect poverty.

Therefore, reliability is a necessary condition for good measurement but not a sufficient one.

We need to make sure that our indicators are effectively capturing poverty.

Validity

- Imagine that we know the standards of living of two subjects in a sample- one highly educated, wealthy and healthy and another with low education attainment, with a lot of debt and with systematic health problems.
- However, we find an unexpected result. The first subject is ranked lower than the second one, i.e. she is more likely to be poor than the second subject.
- Measurement theory tells us that our scale is reliable but invalid. That means that there is very little evidence to interpret our index in accordance with our theory and concept of poverty.

Validity

- A valid measure is one that tells use the nature of what is being measured and its relationship with the index in question to its cause.
- Validity is a property that aims to assess the extent to which an index captures what we mean to measure.
- In other fields, one could ask someone the amount of sugary drinks (in ml) they had in a week. This information could be recoded using a questionnaire, for example. How can we validate this measurement? Well, we could follow someone everywhere and every time and take notes of their drinking consumption. Then we could compare our measurement to hers to assess the precision of our instrument.

Validity with unobserved constructs

Can we follow the same strategy in poverty research?

- No, we cannot as we work with an unobserved construct.

The history of the Standards for Educational and Psychological Testing summarises the conceptualisation of validation of constructs (AERA, APA and NCME 2014).

A framework of validity in measurement

- Classical test theory (CTT) proposes that reliability is the maximum possible validity of a scale.
- Validity is a function of systematic error and results in deviations from the construct of interest.
- That means that a scale can be reliable but always wrong because it always deviates from the target of interest.

The different types of validity

- Bandalos (2018) provides and overview of how the definition of validity has changed over time.
- In the 1950s, criteria and predictive validity were the dominant approaches in both psychometrics and educational measurement literature.
- These two forms of validity focused on the correlation between the scale in question and a predictor of the phenomenon of interest.
- In our example, criterion validity would have shown that our scale had an inverse relationship with some observable attributes of the subject in the sample.

Therefore, the scale would have been regarded as invalid from the perspective of criterion validity.

Criterion validity

- Criterion validity demands a clear theory about the causes and consequences of the phenomenon of interest.
- Townsend (1979) provides a good framework for such a purpose in that it provides a clear causal mechanism: command of resources, poverty and deprivation.

Therefore, measures of command of resources (another latent construct) could be used to predict poverty.

Drawing upon, Townsend (1979), criterion validity has been used in poverty measurement by Guio, Gordon, and Marlier (2012) for the production of the European deprivation index and by Gordon (2010) in his proposal for the Mexican multidimensional measure. Similarly, Nandy and Pomati (2015) used criterion validity to assess their proposed index for Benin.

Content validity

The association of an index with a predictive criterion may be inadequate or infeasible in some circumstances.

In practice, some scales are developed to target certain aspects of a construct.

- Policymakers or institutions might prioritise some aspects of poverty from a human rights perspective, for example. This consideration leads to content validity.
- In poverty measurement, perhaps the most emblematic recent example is the Mexican measure. Drawing upon the Mexican Constitution (1917), the National Social Development Law defined poverty in terms of social rights.

Face validity

- One critical question about content validity is about how does a researcher **knows** or **defines** the constituent parts of the phenomenon of interests.
- Most of the time these aspects come from theory. However, some experts tend to flight on first class all the time. Do they know something about poverty?
- The use of mixed methods is a way to enhance the capacity of theorists to develop concepts and frameworks about the mechanisms through which such concepts interact.
- **Face validity** is a form of validation that comes mainly from qualitative work. One way to see face validity is thinking in terms of how transparent a test looks like for the participants of the measurement.
- There are several qualitative methods to assess face validity and the best implementation to date is the Poverty and Social Exclusion project implementation of the Consensual Method (Pantazis, Gordon, and Levitas (2006); Gordon (2018)).

Construct validity

Cronbach and Meehl (1955) put forward a third form of validity that suggest that the measurement of the construct should be useful to *meaningfully* split groups.

Whereas reliability guarantees certain ordering, construct validity focuses on the meaning of such ranking.

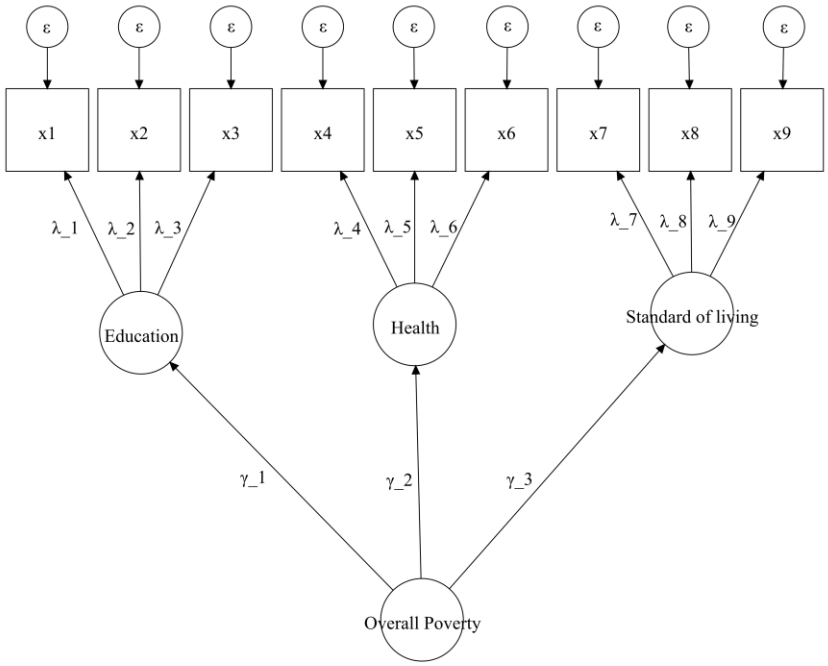
- Construct validity requires mounting evidence in favour that the scale does what is meant to do. Messick (1987) argued that construct validity embraces almost all types of validity evidence.

For him, all the available evidence on a scale adds to the latent rejection or continuity of a scale. AERA, APA and NCME (2014) define validity as (p.14):

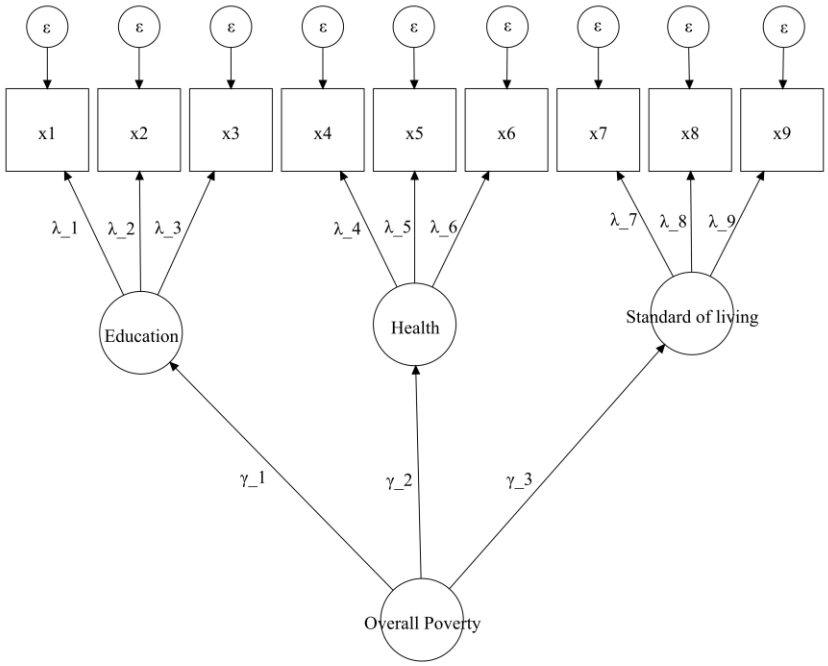
> *It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed used.*

This modern definition refers thus to the different types of evidence on the validity of a scale- criterion, predictive and content.

Construct validity aims to say whether this model is



This is a visual representation of Alkire and Santos (2010)'s model. Second-order factor



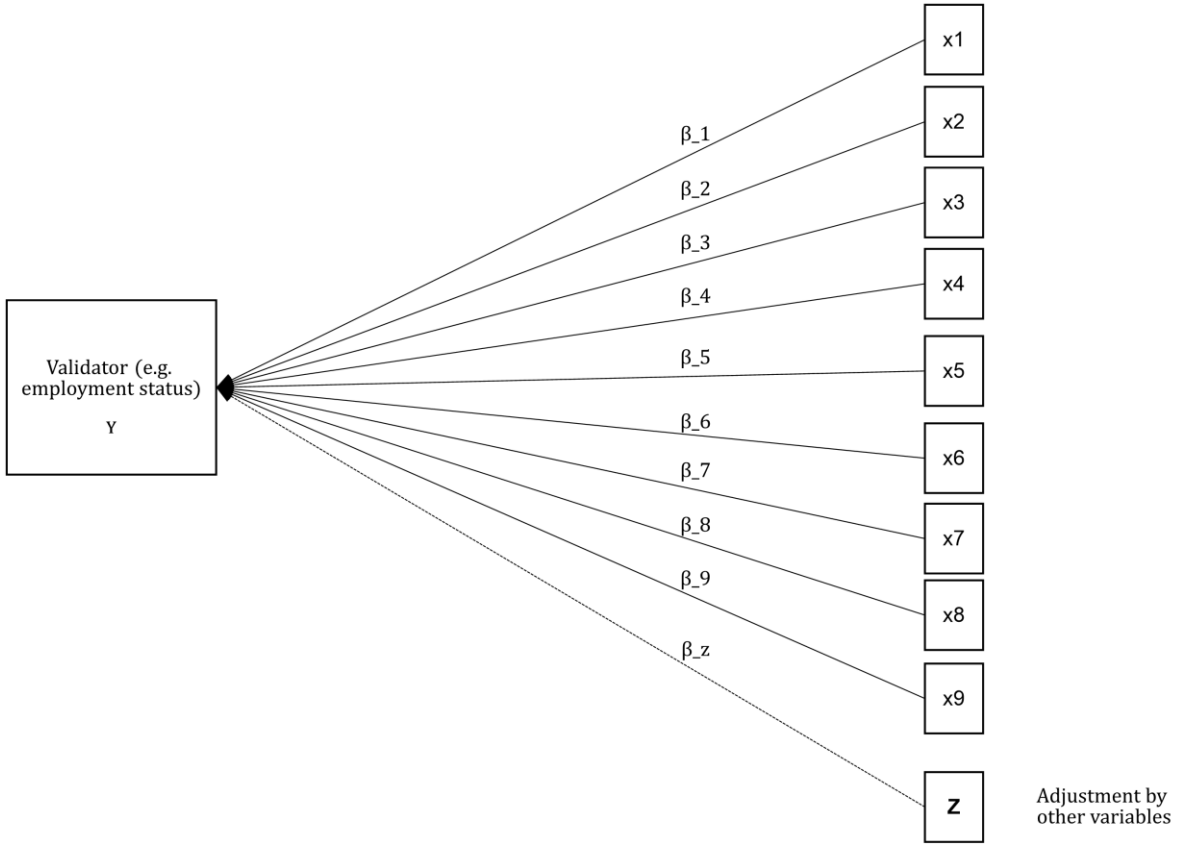
This is a visual representation of Alkire and Santos (2010)'s model. Second-order factor

Methods for scale validation

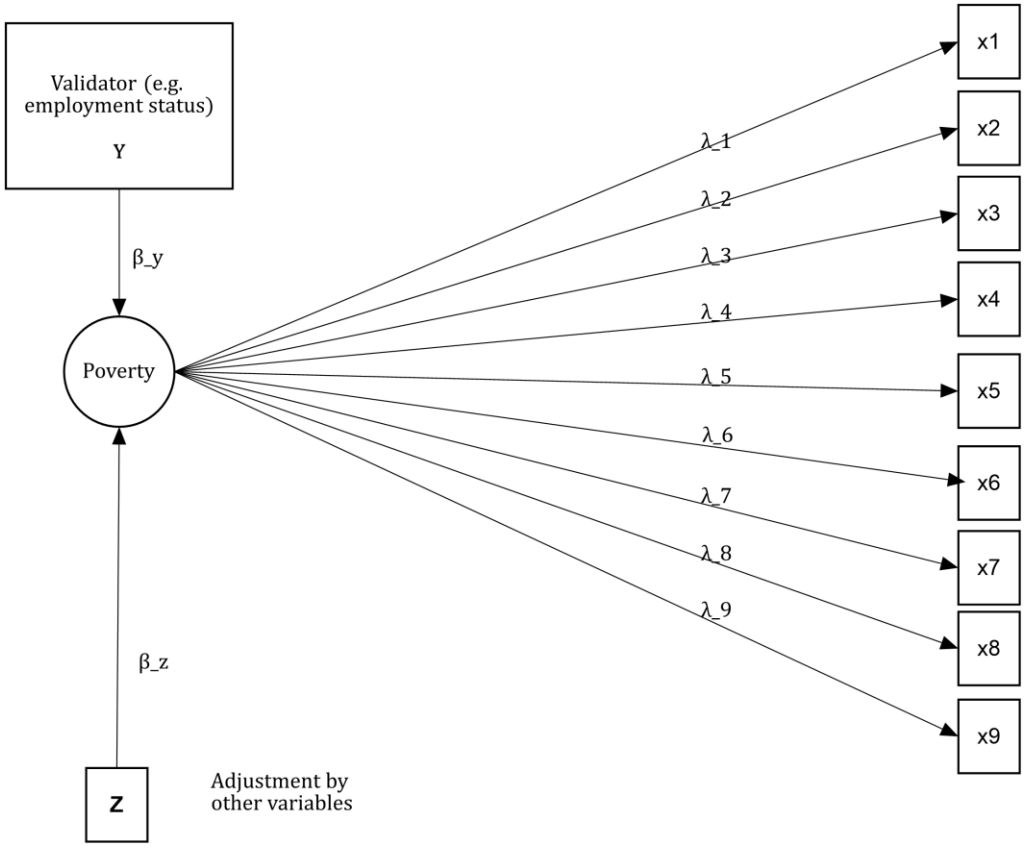
Criterion validity

Criterion validation is characterised by the correlation between an index and an alternative measure on the cause or effects of the construct of interest.

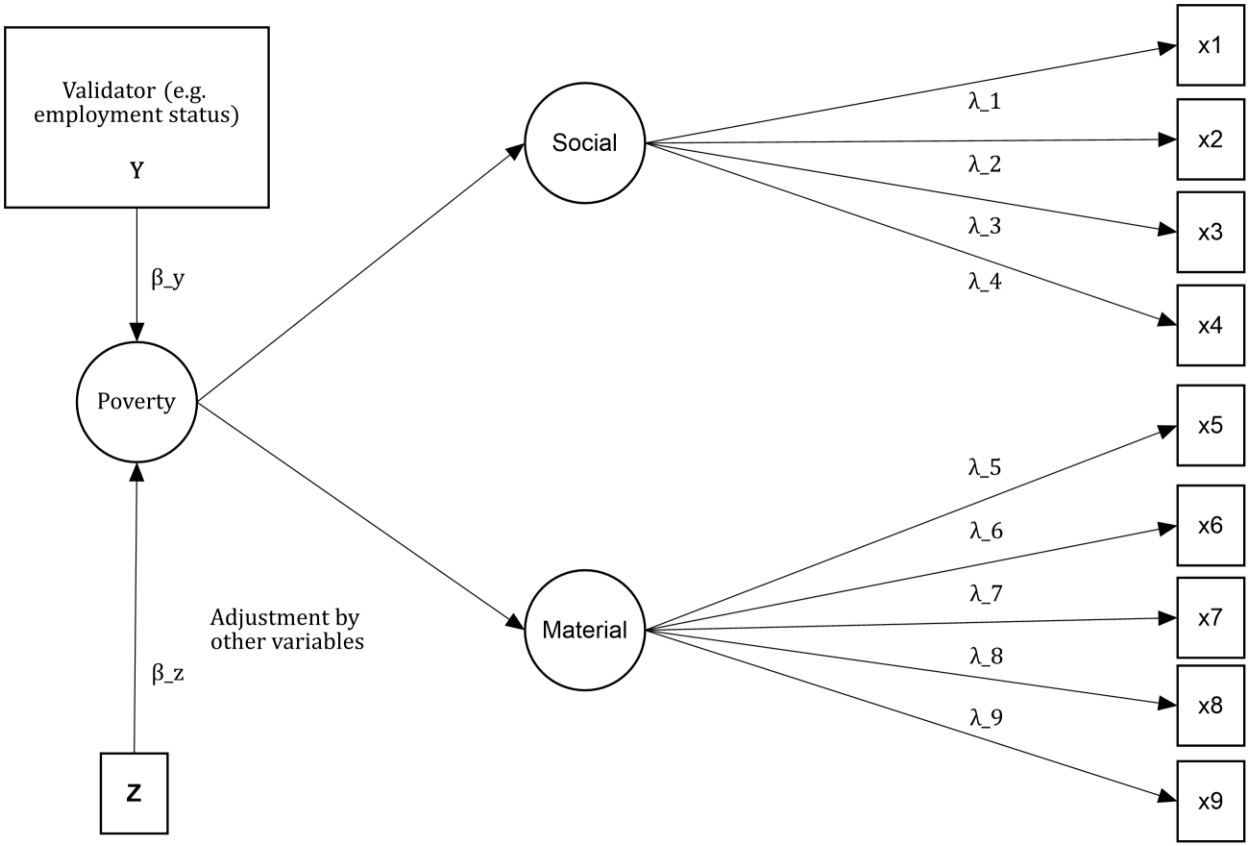
- This requires a theoretical framework explaining the drivers and consequences of poverty as well as how these two relate with the concept of deprivation. There are several good books on theories of poverty (see for an overview of different sources)(Spicker, Alvarez, and Gordon 2006)
- Gordon (2010) proposes fitting a regression model to assess the extent to which the (reliable) indicators of a poverty measure correlate with a proxy measure of command of resources.
- Income poverty (Poor=1 and Not poor=2) as a response variable and the deprivation indicators as predictors. The model was adjusted by urban/rural and household size. The expectation thus was to find relative risks ratios higher than 1 ($\beta_i > 1$) as this is an indication that being deprived of a given item increased the chances of being classified as poor.



This is a visual representation of Gordon (2010) criterion validation. Here income poverty is replaced by a measure of socio-economic position like employment status.



This is a visual representation of a MIMIC criterion validation of a unidimensional or null model



This is a visual representation of a MIMIC criterion validation of a reduced version of the theoretical model of Townsend

Construct validity

- Construct validity is an ongoing process and it is part of a unified framework of validity.
- Model specification is central in a statistical framework to measure poverty. This entails making explicit assumptions about the number, type and nature of the dimensions and its indicators.

It also involves making assumptions about how the model should behave, i.e. people with multiple deprivation should be more deprived than people with a single or no deprivations, for example.

- Construct validity comprises different sorts of evidence on the different hypothesis of the measurement model.

Construct validity

To illustrate this we will use the Multidimensional Poverty Measure of acute poverty.

- Multidimensional poverty has three substantive dimensions: education, health and standard of living.
- These dimensions are clearly distinguishable (discriminant validity).
- The indicators of each dimensions are adequate manifestations of deprivation of education, health and standard of living (classification of indicators).
- The indicators of each dimensions equally account by for variation of the sub-dimensions (within-dimension weights).
- The four hypothesis underpin the measurement model of poverty of the MPI. These are ordered from the more general to the most specific.
- How then these assumptions could be tested. Measurement theory has developed factor models for such a purpose.

University of BRISTOL

Universidad Nacional Autónoma de México UNAM

Construct validity and Confirmatory Factor Models

Measurement models have a series of parameters (item loadings, dimension loadings, item thresholds and errors).

Confirmatory factor analysis (CFA) is a way to estimate the value of the parameters in question and assess the extent to which the model reproduces the observable relationships among the indicators.

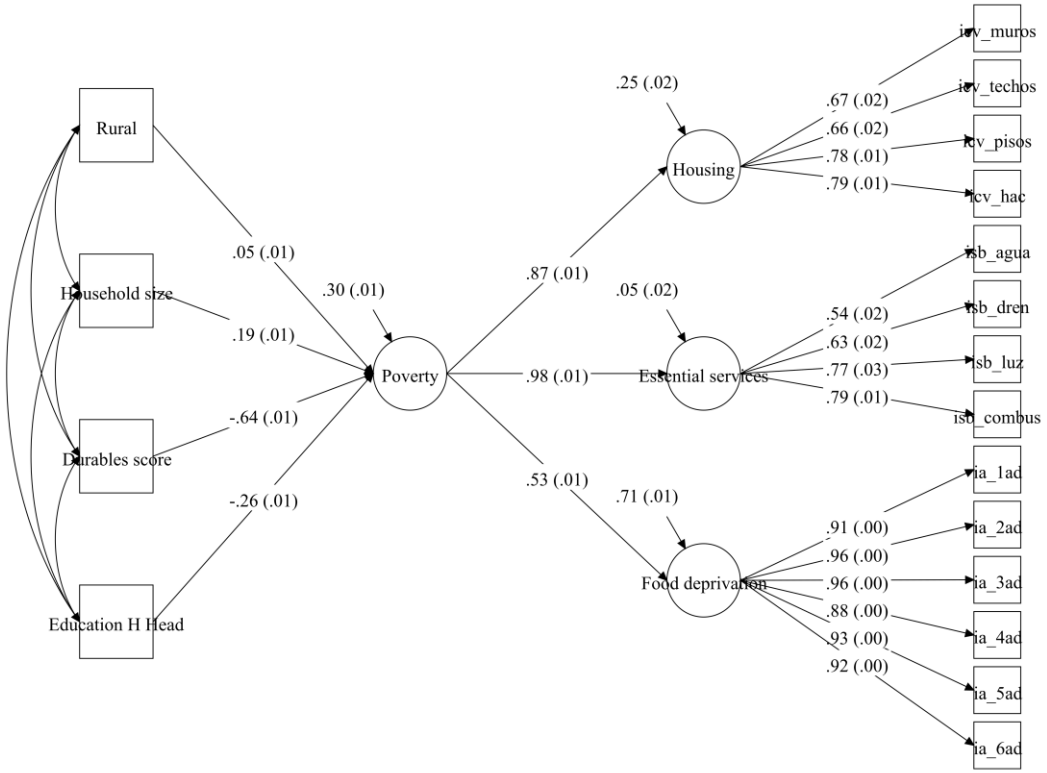
How does CFA assesses whether a model matches observation?

CFA estimates a series of parameters that produce a variance-covariance matrix (Σ) that approximates as closely as possible the observed variance-covariance matrix (S). Therefore, the goal in CFA is to find a set of parameters that best reproduces the input matrix. This process is achieved by minimizing the difference between Σ and S.

CFA and overall fit

- F_{ML} is used for several goodness-of-fit indices. An absolute index is χ^2 which operates with the null hypothesis that $S = \Sigma$. When rejected, it tells that the proposed model is not good enough to reproduce S. In other words, the number, type of dimensions and indicators do not result in an adequate representation of the construct. $\chi^2 = F_{ML}(N-1)$ and thus is sensible to sample size and based on a very stringent hypothesis that $S = \Sigma$.
- Comparative fit indices use a baseline model (typically a null model) as reference to evaluate the fit of the proposed model. These indexes often look more favourable than the strict χ^2.
- Extensive Monte Carlo studies have found that these indexes are nonetheless trustworthy and well-behaved.
- The Comparative Fit Index (CFI) is one of the most widely used. It varies between 0 and 1 where values closer to 1 indicate a good model fit.
- The Tucker-Lewis index (TLI) is another popular alternative which includes a penalty function for adding more parameters that do not necessarily improve the fit of the model. It typically has values between 0 and 1, where again closer to 1 implies a relatively good model fit. $> .95$ is considered as good fit.

For example. The mexican measure has good fit (TLI>.95)



This is a MIMIC model of a reduced version of the multidimensional Mexican measure. The model shows that poverty is associated by posession of different goods and education attainment of the household head, adjusted by rurality and household size. Standardised coefficients (Standard error within brackets)

AERA, APA and NCME. 2014. "Standards for Educational and Psychological Testing." Edited by American Educational Research (AERA) and American Psychological Association (APA) and National Council on Measurement in Education (NCME) and Joint Committee on Standards for Educational and Psychological Testing (US). Amer Educational Research Assn.

Alkire, S., and M. Santos. 2010. "Acute Multidimensional Poverty: A New Index for Developing Countries." OPHI Working Paper No. 38.

Bandalos, Deborah L. 2018. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications.

Brown, T. 2006. *Confirmatory Factor Analysis for Applied Research*. Edited by T Brown. The Guilford Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16: 297–334.

Decancq, Koen, and María Ana Lugo. 2013. "Weights in Multidimensional Indices of Wellbeing: An Overview." *Econometric Reviews* 32 (1): 7–34. https://doi.org/10.1080/07474938.2012.690641.

Gibbons, Robert D., Jason C. Immekus, R. Darrell Bock, and Robert D. Gibbons. 2007. "The Added Value of Multidimensional Irt Models." *Multidimensional and Hierarchical Modeling Monograph* 1.

Gordon, D. 2010. "Metodología de Medición Multidimensional de La Pobreza a Partir Del Concepto de Privación Relativa." In *La Medicion de La Pobreza Multidimensional En México*, edited by M. Mora, 401–98. El Colegio de México. CONEVAL.

———. 2018. "Measuring Poverty in the Uk." In *Poverty and Social Exclusion in the Uk*, edited by E Dermott and Gill Main.

Guio, A., D. Gordon, and E. Marlier. 2012. "MEASURING Material Deprivation in the Eu: Indicators for the Whole Population and Child-Specific Indicators." EUROSTAT.

Guio, Anne-Catherine, Eric Marlier, David Gordon, Eldin Fahmy, Shailen Nandy, and Marco Pomati. 2016. "Improving the Measurement of Material Deprivation at the European Union Level." *Journal of European Social Policy* 26 (3): 219–333. https://doi.org/10.1177/0958928716642947.

Guttman, Louis. 1945. "A Basis for Analyzing Test-Retest Reliability." *Psychometrika* 10 (4): 255–82. https://doi.org/10.1007/BF02288892.

Harris, Deborah. 1989. "Comparison of 1, 2, and 3-Parameter Irt Models." *Educational Measurement: Issues and Practice* 8 (1): 35–41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x.

Loken, Eric, and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355 (6325): 584–85. https://doi.org/10.1126/science.aal3618.

McDonald, R. P. 1999. *Test Theory: A Unified Treatment*. Edited by R. P. McDonald. Mahwah, N.J. L. Erlbaum Associates.

Messick, Samuel. 1987. "Validity." *ETS Research Report Series* 1987 (2): i–208.

Muthén, Beng. 2013. "IRT in Mplus." Mplus. http://www.statmodel.com/download/MplusIRT2.pdf.

Nandy, Shailen, and Marco Pomati. 2015. "Applying the Consensual Method of Estimating Poverty in a Low Income African Setting." *Social Indicators Research* 124 (3): 693–726. https://doi.org/10.1007/s11205-014-0819-z.

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

Pantazis, C., D. Gordon, and R. Levitas. 2006. *Poverty and Social Exclusion in Britain: The Millennium Survey*. Edited by C. Pantazis, D. Gordon, and R. Levitas. Studies in Poverty, Inequality, and Social Exclusion. Policy Press. https://books.google.com/books?id=o-H0J4BMWS8C.

Revelle, William, and Richard. Zinbarg. 2009. "Coefficients Alpha, Beta, Omega, and the Glb: Comments on Sijtsma." *Psychometrika* 74 (1): 145–54. https://doi.org/10.1007/s11336-008-9102-z.

Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things." *Américan Journal of Psychology* 15 (1): 72–101.

Spicker, P., S. Alvarez, and D. Gordon. 2006. *Poverty and International Glossary*. Edited by P. Spicker, Alvarez S., and D. Gordon. International Studies in Poverty Research. International Social Science Council. Zen Books.

Streiner, David L., Geoffrey R. Norman, and John Cairney. 2015. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press, USA.

Townsend, P. 1979. *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living*. Edited by P. Townsend. University of California.

Zinbarg, RichardE., William Revelle, Iftah Yovel, and Wen Li. 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): 123–33. https://doi.org/10.1007/s11336-003-0974-7.