Multidimensional poverty measurement: A statistical approach with applications

Héctor Nájera

Contents

Preface

What is the extent of poverty? Why some population groups or regions are more likely to be poor than others? These answers two these questions are decisive for social policies in that they shed light upon the state of fairness and social justice in society. Accurate and precise answers are crucial because high uncertainty numbs and deviates our reasoning and judgement about how many people are poor and why.

The international consensus is that poverty is multidimensional and that it should be measured taking the substantive aspects of people's necessities of life. However, the answer about the extent and nature of multidimensional poverty is contested and unsatisfactory. There are several theoretical and methodological reasons impeding the production of uncontested poverty measures. There are several theories of human needs that debate the substantive aspects that humans must have to live with dignity. Furthermore, even if these theoretical discussions find a satisfactory solution, several practical obstacles require to be addressed for the successful development of poverty scales. These challenges occur at different stages of the production of an index starting with the fact that poverty is a human invention that needs to be tractable using multivariate data to accurately capture its substantive aspects. This turns out to be a noise-magnifying process in that researchers make several assumptions about the relevant set of necessities to include in an index, the thresholds to identify deprivation, the weighting scheme to reflect the importance of different needs, the way in which the poor and not poor are identified. Furthermore, all these assumptions are constrained by the available data, which is seldom collected with the a priori idea of measuring poverty. Invariably, researchers have to make decisions and presumptions which are influenced by biases (plus random error). Therefore, poverty measurement requires a cogent framework to mitigate confirmation biases by putting our assumptions to scrutiny.

This book focuses on one of the key problems in contemporary poverty measurement- the lack of a framework to assess the assumptions underlying a multidimensional scale. The book provides a series of tools to fight against prejudice, misconception and error in poverty measurement. It draws upon measurement theory, with a history of 100 years of continuous development, to help researchers to avoid producing noise-magnifying indices. This book provides a series of falsifiable principles and criteria to assess whether our poverty rates are just a reflection of noise mining and unlikely to replicate. Poverty research has overlooked the developments in other fields and although sometimes some aspects of measurement theory are recovered its use is partial, inaccurate and unsystematic. In 2016, the World Bank Commission (2017) on Global Poverty, headed by Sir Anthony Atkinson, has put into perspective the different challenges in multidimensional poverty measurement and set out 21 recommendations. Recommendation 4 of the World Bank report acknowledges the need of validating poverty indices. But it does not propose how to do so. This is understandable because one of the one of the main difficulties in contemporary poverty measurement is the absence of an explicit discussion about how to check all these assumptions.

The most common practice still consists in using ad hoc or idiosyncratic methods to assess some arbitrary properties of a multidimensional scale. To date very few exercises in multidimensional poverty measurement rely on an explicit statistical framework to put under scrutiny the assumptions of an index. Guio, Gordon, & Marlier (2012) and Guio, Gordon, Marlier, Najera, & Pomati (2017) is perhaps the most comprehensive implementation that draws upon the experience from the Poverty and Social Exclusion (PSE) and Peter Townsend's outstanding work. However, the methods its application using software are not widely available for the community interested in measuring poverty. One of the problems is that there are several books

on poverty measurement but there is none exclusively dedicated to the topic of empirical examination of poverty indices. Most of the books focus on the production of an index but devote little attention to the issue of validation. This reflects the fact that poverty research has followed the path of many other fields. There have been other disciplines in a similar struggle. Educational testing, psychological measurement, sociology but also in the natural sciences biology and medicine often face measurement challenges. However, many of these disciplines have taken measurement very seriously and have adopted a series of practices and principles that reduce uncertainty about the attribute they measure. These areas rely on the seminal work of (Spearman, 1904) on correlation and latent variables that resulted in the development measurement theory and methods with more than 100 years of history and continuous development from the classical works on factor analysis (Cudeck & MacCallum, 2012; Lazardfeld & Henry, 1968; Thorndike & Hagen, 1969; Thurstone, 1947), passing through the development of the principles of validity and reliability (Guttman, 1945, pp. Novick1967, Novick1967, Brennan2006), then through the modern framework of the latent variable approach (Bartholomew, 1987; Kvalheim, 2012; Muthén, 2007; Skrondal & Rabe-Hesketh, 2007) and finally to the classic handbooks that show how all these principles and method constitute sound measurement framework (Allen & Yen, 2001; Brennan, 2006; McDonald, 2013; Michell, 2015; Streiner, Norman, & Cairney, 2015). This framework has been so widely accepted that has lead to the adoption of standards in some academic journals so that authors provide a more objective judgement about the quality of their measurement.

The book draws upon measurement theory and methods that have proven to be useful in many other fields, to illustrate how a unified framework can be for empirical examination of multidimensional poverty measures. It translates key concepts and principles of measurement theory and methods and illustrate is implementation using both simulated and real data examples. The book is intended for applied researchers and students. Most of the examples rely on **R-software** and **Mplus** (Muthén & Muthén, 2012; R Core Team, 2018).

The principal goal of this book is to help researchers, students and technicians at the government to understand the importance of the principles of measurement theory in poverty measurement and to enhance their skills for empirical analyses of poverty indices. After studying the book readers should be able to:

- Understand why is important to have falsifiable measures in poverty research
- Identify the difference between a method of aggregation and a methodology for empirical examination
- Appreciate the relevance of measurement theory to examine poverty indices but also to understand its limitations
- Understand how the principles of reliability and validity are a necessary condition for a minimum quality of measurement
- Implement analysis of reliability and validity in widely used software
- Interpret the results of the analysis critically
- Appreciate the role measurement invariance and scale equating for the comparison of poverty indices
- Implement basic analysis of measurement invariance and equating in poverty measurement
- Identify appropriate and inappropriate uses of the method and principles of measurement theory

The book is organised as follows. The first chapter introduces the links between the problems in poverty measurement and the principles of measurement theory. The chapter starts by overweening some of the key debates and consensuses in the literature. It puts emphasis on the challenges and assumptions that take place when measuring poverty and the possible response from measurement theory. The second chapter introduces, discusses and translates the concept of reliability to poverty measurement. It uses simulated data and real data to illustrate the consequences of violating reliability and shows how reliability is deeply connected with some axioms in poverty measurement. The third chapter presents the concept of validity and relates the different types of validity to the checks that can be done in poverty research. Chapter four concerns with the topic of comparability in poverty measurement. It shows how the principle of measurement invariance is central for making valid comparisons across groups and periods. Chapter five continues with the topic of comparability but focuses on the issue of making seemingly incomparable scales comparable. It draws on the principles and method of scale linking and equating.

This book would have been possible without the support...

Chapter 1

Poverty and measurement theory principles

Abstract

This chapter introduces the concept of poverty and draws upon measurement theory to frame some of the challenges in the production and empirical assessment of poverty indices. The roles of poverty definitions, researcher's value judgements, desirable properties of poverty indices, survey data and measurement error in the production of poverty measures are described.

1.1 The Concept of Poverty

Poverty is one of the capital concepts in social sciences. As such, poverty is a construction of the human mind to depict a state of low living standards. In the realm of social sciences, concepts or constructs are not attribute or feature directly observable using univariate data. Yet, poverty is something that can be grasped intuitively by anyone and, at the same time, a construct with several contested interpretations. This is why is so difficult to measure it because there is a subtle but decisive distinction between direct and indirect observation of a given construct.

Poverty has several meanings and part of the difficulty in measuring it has to do with the existence of different definitions. (Spicker, Alvarez, & Gordon, 2006) suggest that the definitions proposed in the literature can be clustered into three main groups: material (needs, resources and deprivation), economic (living standards, inequality and economic position) and social conditions (entitlements, social security, exclusion, dependence and social class). Spicker et al. (2006), nonetheless, underline the importance of working with scientific definitions of poverty as they meet the standards of the philosophy of science: definitions that are testable so that are falsifiable in a clear way. The contemporary literature the poor are defined in terms of both low living standards and resources. This broad definition suggests the existence of two sub-population groups that are meaningful and falsifiable in the sense that their profile should predict outcomes that in theory are caused by poverty- mortality, poor health, economic stress, etc. The chief objective of this books is to provide a framework to make poverty indices falsifiable.

(Townsend, 1979) argued that poverty can be treated scientifically in that it can be objectively defined and measured. According to his theory, the concept of *deprivation* was central for the definition of poverty in that it connects command of resources with low living standards. Poverty can be defined as the lack of resources overtime where material and social deprivations are its consequences (Gordon, 2006). This definition does not clarify in what sense lacking something is a standard to classify people as poor or not poor.

This is related to the domain that is utilized to identify deprivation and in the literature has to do with the discussion about absolute and relative poverty. Townsend (1987) argued that poverty is relative in the sense that it varies across time and space- the identification of the relevant domain depends on what a society

regards as the minimum according to the prevailing living standards. (Sen, 1983) was the most notable thinker against to the idea of poverty as a relative concept. He suggested that poverty was absolute in the sense of not having certain basic opportunities- failure in capabilities. (Altimir, 1979) concluded that Sen and Townsend were talking about two nested thresholds. The absolute core which is universal and a relative one which varies across societies and time.

In a series of exchanges in Oxford Economic Papers in the 1980s, Townsend and Sen discussed their views on poverty as an absolute or relative concept. As (Gordon, 2006) points out most of the disagreement is a matter of semantics. (Boltvinik, 1998) provides an recapitulation of the exchanges and he notes that part of the disagreement has to do with the lack of clarity on the space to identify poverty: commodities, resources, capabilities and deprivation. Sen (1983) acknowledged that commodities and characteristics change overtime but to identify poverty researchers must establish when people fail to achieve certain minimum capabilities. There are however, two difficulties in operationalising capabilities. Firts, Thorbecke (2007) points out that measuring capabilities implies observing them ex ante but in practice only outcomes -achieved functionings- can be measured. Second, the is the challenge of specifying the minimum capability set. Whereas some authors have proposed a minimum list (Nussbaum, 2000); others like Sen (2005) and Alkire (2007) have been against this idea. At the core of this debate seems to be a lack of clear distinction between a theoretical list (which can be authoritative if imposed without any sort of evidence) and a backed-up list by some sort of validation via public reasoning or empirical exercises.

Measurement of achieved functionings provides the basis to resolve the dispute about absolute-relative definitions of poverty (Spicker et al., 2006). As it implies that capabilities are operationalised through socially defined commodities and characteristics (the next section discusses how this fits a measurement framework that draws on latent variables).[1] In the space of outcomes (X), there is very little practical difference between the concepts of achievements and deprivation x. However, the question about what are the contents (X) of the definition of poverty remains unanswered at this point. The solution lies in establishing the space of functionings (that relate to deprivation capability in Sen's terms) or the space of deprivation according to the standards of society in Townsend's framework. Both authors acknowledged that such space is multidimensional in the sense that it relates to the minimum diverse aspects than enable humans to function/participate in society.

1.2 Theoretical dimensions of poverty

Theories of human need, capabilities and relative deprivation have been put forward to frame the (j) dimensions and its contents (x_{ij}) that should be included in both the definition and measure of poverty. Most notably, the Unsatisfied Basic Needs approach, with a long track record in Latin America, draws on theories of human need such as those proposed by Maslow (1943) and Max-Neef, Elizalde, & Hopenhayn (1992). Altimir (1979) and Boltvinik & Hernández-Láos (2001) draw upon the concept of human needs (instead of capability or relative deprivation) to define poverty in terms of unmet basic needs. The UBN approach has had at its core the housing dimension (access to water and sanitation and materials of the dwelling) plus education, food and health deprivation. Boltvinik (2014) reviews the different variants of the UBN and it is possible to appreciate the different dimensions of poverty (as well as diverse aggregation methods and strategies to identify the poor) where he identifies as the improved variant the one that includes time and underpins the Integrated Poverty Measurement Method (IPMM) (Boltvinik, 1992, p. Boltvinik2001). A recently popular variant of the UBN are the hybrid approaches that combine UBN and social rights. Notably, UNICEF's first international measure of child poverty proposes eight dimensions (Gordon, Nandy, Pantazis, Pemberton, & Townsend, 2003). A similar hybrid variant of the UBN is the Mexican Multidimensional Measure. It has two domains: income and social rights. The human rights domain has five dimensions: Housing, social security and health, education, essential services and food deprivation (Cortés, 2014, p. CONEVAL2011d). The capability-based dimensional models are very similar to the UBN approaches. For example, drawing upon the capability approach, Klasen (2000) proposes a core deprivation index that it is fairly similar to the standard UBN dimensions. This is understandable as it focuses on outcomes that often relate to basic

[1] The notation is fully introduced in the following sections but will introduce it also little by little to help not mathematical readers: specific outcomes/deprivation/achievements $\mathbf{x}$ and the full set of outcomes $\mathbf{X}$)

human needs. These models are fairly recent and its dimensional structure heavily draw upon the original UBN variants implemented in the 1970s and 1980s in Latin America (Boltvinik, 2014) (Chapter XX discusses some of the aggregation novelties introduced by this approach). The most popular implementation is the UNDP-OPHI international model for acute poverty which classifies the indicators into 3 dimensions: standard of living, education and health (UNDP, 2014). As in the UBN measures the housing facilities and conditions are central for the measure. The implementations in Latin America have put forward a five-dimensional structure: housing, basic services, living standard, education and employment (Santos & Villatoro, 2016).

Relative deprivation has also a hierarchical structure but it proposes different dimensions and subdimensions. Townsend (1979)'s original model suggested two main domains: material and social deprivation but with four and seven subdimensions, respectively. For material: dietary, clothing, fuel and light, household facilities and amenities, working conditions, health and educational. For social: Environmental, Family, Recreational and Social (Townsend, 1979, pp. 1173–1174). There have been several models that draw upon relative deprivation but the theory behind them is not as explicit as in the case of the UBN or the Townsend model (Betti, Gagliardi, Lemmi, & Verma, 2015; A.-C. Guio, 2009; Whelan, Nolan, & Maitre, 2006). These nonetheless suggest rather different dimensional structure. A.-C. Guio (2009) proposes a three dimensional model comprising economic strain, enforced lack of durables and housing-related deprivation; Whelan et al. (2006) propose five dimensions (economic strain, consumption, housing facilities, neighbourhood environment, health status) and Betti et al. (2015) put forward seven dimensions that are contain most of Whelan et al. (2006)'s proposal (Chapter X discusses more in detail the implications of these models).

Drawing upon this review of the conceptualization of poverty, throughout the book we will be referring to Townsend (1979) definition (Gordon, 2006):

> *Poverty is the lack of command of resources overtime and deprivations its consequence.*

This concept of poverty is useful because it is simple and is not in tension with many of the conceptual debates described above. The definition acknowledges that Deprivation is multidimensional in that it refers to the unmet needs/functionings that are regarded as essential by society in a given point in time. But, the definition is also useful

Therefore, once the set of socially defined needs is established, the question is how the poor is correctly identified? This is the question of this book: What criteria and principles lead to falsify the contents and identification of a poverty scale? To better frame the question is appropriate to review the challenges involved in poverty measurement.

1.3 The measurement of poverty and its challenges

The concept of poverty involves raising a series of assumptions about the existence of a minimum living standard that permits meaningfully to split a population into two groups. Researchers therefore have to make several decisions with regard how to select the dimensions and its indicators but also about how what is the best way to aggregate the information so that the poor is accurately identified (Alkire, 2007; Thorbecke, 2007). These result into a series of decisions that demand theoretical and empirical justification:

1. Specifying the j dimensions
2. Specifying the contents (indicators) of each dimension x_{ij}
3. Deciding the cut off $x_{ij} < z$ of the indicators
4. item Establishing a measurement model about the relationship between dimensions and indicators
5. item Deciding the relative contribution (weights) w_j of some indicators/dimensions
6. item Deciding how to aggregate the information to rank the population
7. item Deciding how to split such ranking into two meaningful groups $p_k(x_i; z) = 1$ if $c_i \leq k$ and $p_k(x_i; z) = 0$ otherwise

1.3.1 Challenges in selection of dimensions, contents, cut offs and weights

With regard the first three, Gordon & Nandy (2012)] point out that although there is a consensus that poverty is multidimensional, there is little consensus about the number, contents and nature of the dimensions of

poverty. Theoretical frameworks rarely go that far and, although there is some overlap of the dimensions used in different perspectives, there is very little agreement about the content and the nature of the interactions between and within dimensions. Alkire (2007) provides and overview of the different approaches that a researcher can employ to decide the dimensions, indicators and its interaction (public consensus, existing data, convention, deliberative participatory process, data on people's values). The typical dimensions proposed from the capability, UBN and relative deprivation were overviewed in the previous section. One critical example, in that is grounded on theory but includes an democratic empirical component is the consensual method pioneered by Mack & Lansley (1985). The consensual deprivation approach was a ground-breaking approach in that it linked the theory of relative deprivation with an actual account of the socially perceived needs of the population. More recently, from the capability perspective, other studies have also aimed at finding the necessities of life by asking the population (Clark & Qizilbash, 2005; Narayan, 2001).

Although the specification via theory and/or some form of empirical data helps to delimit the space of relevant needs, it does not solve problems about the interaction between indicators and dimensions. More importantly, it does not guarantee that such a list would effectively result in an accurate account of the dimensions of poverty. There are several reasons. One criticism is reflected in the argument pose by McKay (2004) about the difference between needs and preferences (i.e. How to split between wishes from constraint?). Another critique is about on what basis one decides what is a need from what is not when there is no 100% endorsement (Bradshaw, Holmes, & Hallerod, 1995). Another argument has been about the difficulties to conduct comparative work, i.e. how to compare countries with different sets of needs. There is, from statistical point of view -which is the topic of this book-, a satisfactory response to these concerns. We will recover these critiques throughout the book as a key task of it is to provide a profound explanation and illustration of why once some principles of measurement theory are fulfilled, these critiques do not hold.

The determination of thresholds levels for the indicators and dimensions is central as it is one of the main sources of variability of an index. Thorbecke (2007) points out that this is often done on a subjective or normative fashion and that it could lead to important discrepancies and reproducibility problems. Aside theory-driven ways to set the thresholds of nominal variables, the consensual approach could be also used to find a split. The best example is the Mexican Multidimensional Measure in that the cut offs where derived from the consensual method and a revision of the norms in Mexico CONEVAL (2011a) (Chapter XX shows how this hybrid approach leads to a more robust measure). Rightly so, Thorbecke (2007) indicates that the key is to find a series of cut offs that yield to a consistent ranking of the population given the observed outcome measures of deprivation.

1.3.2 Challenges in aggregation and identification of the poor

Finding the spaces of needs, dimensions and cut offs is just one stage in poverty measurement. (Thorbecke, 2007, p. 7) makes the following point:

> *Now let us assume that, notwithstanding all the difficulties discussed above, agreement has been reached on a list of attributes related to poverty and their threshold levels. How can such information be used to derive measures of multidimensional poverty and make poverty comparisons? Start with the simplest case, for example, that of an individual who is below each and every attribute threshold level. Such a person would be classified as unambiguously poor. Analogously, comparing two individual poverty profiles (A and B) where the attribute scores for all of the n dimensions in the profile of A are above that of the profile of B, it can be inferred unambiguously that A is better off in terms of well-being (less poor) than B.*

The question about how to aggregate the indicators and then how to split the population into two meaningful groups (i.e. poor and not poor) has been present since the classic studies of poverty. From a theoretical perspective there have been three responses to this question. Townsend (1979) put forward a theory that stated that there must be a level of resources (the cause) from which deprivation raises substantially (consequence) (This is also known as the Townsend breaking point). Such a point should be the poverty line in that it leads to a meaningful split of the population, i.e. people whose standard of living is so low that they are effectively excluded from the patterns of living in society. Within this approach the aggregation consists in counting the number of deprivations and finding an optimal split based on its relationship with resources. P.

Townsend & Gordon (1993) argue that when using cross-sectional data, the best approach is the intersection -below certain level of resources and above certain level of deprivation- (See (Gordon, 2010)). This is the approach, for example, adopted to identify the poor in Mexico (CONEVAL, 2011a). One distinctive aspect of the Townsend breaking point is that the dimensions do not figure although the accounting of the diversity of attributes of deprivation remains.

A second theory-driven approach consists in using the union approach (being deprived in either resources or basic needs). This approach is used as part of Boltvinik's integrated method (Boltvinik & Hernández-Láos, 2001). The IMMP aims to minimize the exclusion of the poor and the view that income, deprivation of needs and time deprivation are three attributes to characterise poverty. This is different from (Townsend, 1979)'s theory where needs deprivation is a manifestation of low command of resources (See (Gordon, 2010, p. @Cortes2014) for a discussion from the perspective of poverty dynamics and human rights).

Unlike the first two approaches, the third main approach focuses more on the aggregation of the indicators than on how to set the poverty line (Foster, Greer, & Thorbecke, 2010). The axiomatic approach was pioneered by Sen (1976) for income-based measures. Sen (1976) put emphasis on the importance of the aggregation stage in poverty measure as it had consequences for the decomposition and analysis beyond the prevalence of poverty. Two axioms were critical for Sen (1976): monotonicity (poverty rise if income falls) and transfer (poverty rise if there is a transfer from the poor to the rich). Later Foster, Greer, & Thorbecke (1984) extended Sen's axioms and this formulation has been developed for multidimensional measures (Tsui, 2002, pp. Alkire2011a, Alkire2015).

The AF index respect several desirable axioms such as monotonicity and other axioms such as symmetry, replication invariance, scale invariance, poverty focus and population subgroup decomposability (Alkire et al., 2015, see for a complete description). It results in an aggregation method that produces and index with properties that are true without proof.

$$P_{AF}(X;z) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} w_j g_{jk}^{\alpha}(k); \alpha \geq 0$$

In their formulation x are achievements in the form of indicators or dimensions from a matrix X and z and the cut-offs for identifying the poor $x_{ij} < z_j$, i, are individuals or households and j the dimensions. A person is deprived in the dimension ($g_{ij}^0 = 1$). Weights are included via w_j, and $\alpha = 0$ is the adjusted headcount ratio and $\alpha = 1$ the adjusted poverty gap.

One crucial aspect that is often and wrongly overlooked when considering the axiomatic approach is the difference between an aggregation method (formula) and a measurement methodology (series of steps). The Alikre-Foster (AF) family of measures impose a series of axioms so that the behaviour of a formula is predictable. The fact that is based on axioms does not make the measurement of poverty based on the AF method true or correct. The axiomatic approach, as any other measure, is based on the key assumption that the indicators (cut offs), weights and dimensions are a sensible account of poverty. Unlike the aggregation used in the counting approach, this aggregation method, explicitly takes into account relative welfare weights (S. Alkire & Foster, 2011). The issue of weighting will be discussed in detail in Chapter 3 but as in the previous case, measurement theory offers a framework to deal with this issue under a falsifiable framework. One of the aims of this book is to clarify why the principles of measurement theory are a necessary condition for the AF to work, whereas the AF is not so for an index that fulfils the core principles of measurement theory.

1.4 The poor and the not poor: The poverty line

From a conceptual basis, the poverty line could be defined as the minimum living standards acceptable in a society at a given point in time. The poverty line is often expressed in either monetary or non-monetary terms (i.e. deprivation weighted count). This is not a book on income poverty measurement but when using an

indirect approach Van den Bosch (2001) identifies the following methods.[2] Peter Townsend (1993) discusses the problems with this type of approach:

- Budget standards: The price of a specific basket of goods and services. This approach was used by Rowntree (1901) and recently most notably by Bradshaw (1993) and Bradshaw et al. (2008).
- Official standards: An arbitrary price or deprivation count is used by an statistical agency to split the population. It is often based on the minimum income support offered by the social security system.
- Food-ratio method: It assumes that living standards can be judge by comparing the proportion of income spend on necessities.
- Relative method: The income threshold is set to a certain percentage of the mean or median income. Some European countries use 60% of the median income, for example.

The introduction of direct approaches to capture poverty has opened up a discussion about whether income and deprivation should be combined to identify the poor (Boltvinik, 1998, 2014; Boltvinik & Hernández-Láos, 2001; Gordon, 2010). In the literature, this is known as the debate between union and intersection approaches. The union approach sees income and deprivation as two measures of the same phenomenon and therefore being below of a certain cut off in either of the two leads to poverty. In contrast, the intersection approach acknowledges that these should be used jointly to identify the poor, i.e. lacking both sufficient income and being multiply deprived is poverty. For Boltvinik & Hernández-Láos (2001) the union minimizes the exclusion of the poor and the intersection maximizes its exclusion. So the first one overestimates poverty and the second is an underestimate. A third option is a partial union approach (S. Alkire & Roche, 2011; Santos & Villatoro, 2016). Income is first dichotomized using $(x; z)$ a cut off. Then, income is included in a score as another deprivation in the AF aggregation method. The partial inclusion is obtained by using differential weights and by setting a poverty line based on a percentage of the total possible deprivation score (Typically 25%).

One practical argument in favour of the intersection approach is given from the perspective of poverty dynamics. When using cross-sectional data is not possible to assess the individual trend in deprivation followed a rise or fall of resources. That is, some households having just experienced a rise in income are expected to be less deprived in the incoming future. However, the snapshot of the cross-sectional survey would put them as poor under the union approach when in reality they are at risk or vulnerable to poverty (Katzman, 2000). Only with high-quality panel data would be possible to identify the truly poor based on the union approach (Gordon, 2006; Halleröd, 1995).

When discussing poverty lines is important to ask what is the underlying theory behind the different approaches. Unfortunately, theories are rarely explicit and imprecise with regard the poverty line. There are three main approaches to set the poverty line:

- Townsend's breaking point: One explicit framework is given by Townsend (1979) and Peter Townsend (1993). One of the predictions of Townsend's theory is that there is a negative relationship between resources and deprivation. Yet, he proposed that such relationship is not linear and that multiple deprivation raises considerably below a certain level of resources.
- Normative approach based on social rights: Social rights are indivisible and interrelated so being deprived of one right captures a situation of denial of basic human needs. CONEVAL (2011a) uses 1+ deprivation and an income below an income poverty line (drawn from a budget standards approach) to identify the poor (Cortés, 2014).
- Normative approach integrated method: Boltvinik & Hernández-Láos (2001) proposes that poverty has three dimensions: time, resources and UBN. People is regarded as poor when failing to meet one of the three dimensions.
- Arbitrary approach: Without any theoretical justification, some researchers set the poverty at some value below 50% of the total weighted sum of deprivations. For example, if there are 10 deprivation and the cut off is 30%, people with three or more deprivations are regarded as poor.

These theoretical and methodological discrepancies lead to considerably differences in the extent of poverty. This is true even when the same indicators are utilized for two measures with different strategies to set the

[2]The calorie-based income poverty line is another popular method, which is based on the presumption that the poor cannot meet their energy requirements -based on a reference basket-

poverty line. This begs the question about how to tackle this challenge from an empirical perspective. This topic is covered in the next chapter of the book.

1.5 A brief on multidimensional poverty measurement

The debates about the definition of poverty and the best way to measure it have, of course, been reflected in different approaches to capture this phenomenon. The diverse views about how to produce a poverty scale have shaped the history and types of poverty measures. Poverty research and measurement has more than 100 years of history. The ground-breaking and now classic studies of Rowntree (1901) `Poverty: A study of Town Life', @Booth1903's`Life and Labour of the People in London' and Townsend (1979)'s 'Poverty in the UK' were possible after years of relentless research that aimed to understand the extent, distribution and nature of poverty. A common feature across these studies is that all used an ex ante developed questionnaire to capture poverty. This is in stark contrast with current practices in poverty research as poverty is seldom measured using explicitly developed questionnaries. There are important differences among these three monumental studies. Perhaps the most fundamental is that Townsend (1979)'s study was fully theory driven and used direct indicators of poverty (deprivation) and not income.

The legacy of these studies has been such that the poverty research agenda dramatically expanded in the late XX and early XXI Centuries. Figure 1.1 synthetizes rather crudely the recent history of multidimensional poverty measurement. The multidimensional approach to poverty has its roots in both Europe and Latin America in 1960s and 1970s. Latin America was at the forefront in multidimensional poverty measurement thanks to the Unsatisfied Basic Needs (UBN) approach which used direct indicators to capture poverty [Altimir (1979);Beccaria1985;Boltvinik1992;Boltvinik2001}. The indicators have been mainly focused on housing and essential services given that the implementation of the UBN has been constrained by the available data. Boltvinik & Hernández-Láos (2001) adjusted variant of the UBN approach is perhaps the most theoretically comprehensive and progressive in that includes time as a dimension. The UBN has become a widely known tradition in poverty measurement and has shaped contemporary measures such as the United Nations Development Programme (UNPD) multidimensional poverty index (MPI) and has greatly influenced official measures such as the Mexican measure CONEVAL (2011b).

In the developed world, Townsend (1979)'s theory of deprivation not only served to produce the first survey questionnaire to measure multidimensional poverty but also proposed a direct and multidimensional measure of poverty. In the 1980s, the use of direct indicators to measure poverty was further suggested during the series of exchanges between Townsend (1985) and Sen (1983) and Sen (1985). This despite the focus of their argument was relative versus absolute poverty. This is not a book on monetary poverty but Sen's axiomatic framework set the bases for the development of axioms for multidimensional measures Sen (1976) and Foster et al. (1984).

In the late 1980s and 1990s Townsend's influence on poverty measurement was boosted by Mack & Lansley (1985)'s consensual approach that added the exploration of needs as a vital component to relative deprivation theory. Europe continued with the relative deprivation tradition but other parts of the world introduced monetary measurement of poverty (including Latin America), which became the mainstream approach in developing countries and led to the World Bank approach and the inevitable and rich debate about it (Pogge, 2005; Ravallion, 2010; Reddy & Pogge, 2010).

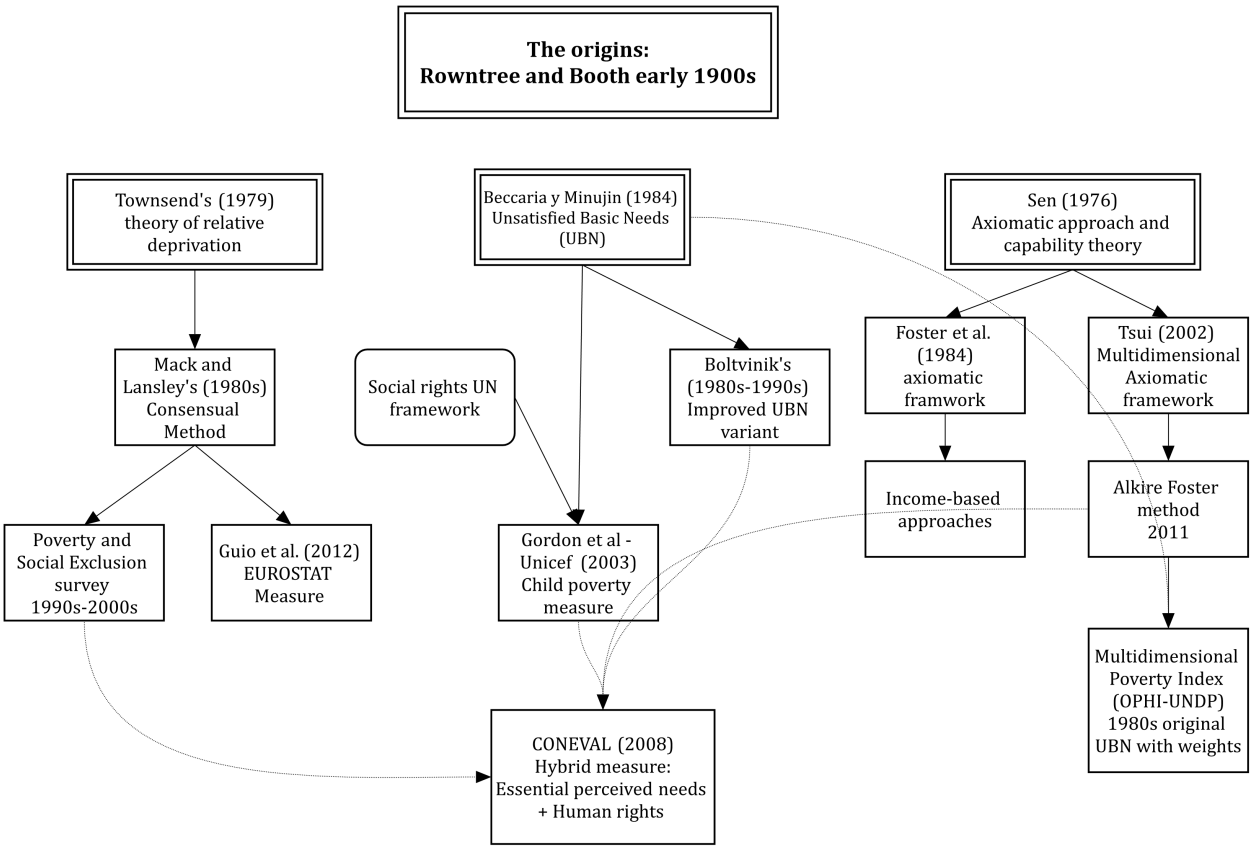

Figure 1.1: A brief summary of the history of poverty measurement

The XXI Century was witnessed the resurgence of multidimensional measurement. In Europe the Poverty and Social Exclusion series has continued Townsend's tradition in the UK Townsend & Gordon (2000), Pantazis, Gordon, & Levitas (2006) and Mack & Lansley (1985) have had a major influence in the measurement of poverty in Europe (Atkinson, Guio, & Marlier, 2017; Whelan et al., 2006). In economics, the axiomatic approach has been finally extended for multidimensional measures (Foster et al., 2010; Tsui, 2002). These contributions and the capability theory as overarching framework has been very influential in multidimensional poverty measurement (S. Alkire & Foster, 2011; Alkire et al., 2015; Kakwani & Silber, 2008). The efforts to articulate human rights and multidimensional poverty measurement have helped to the progressive institutionalisation either by nations or by international institutions of official multidimensional measures (Boltvinik, 2014; CONEVAL, 2011a). The work Gordon et al. (2003) on child poverty pioneered the era of global harmonized multidimensional poverty measurement by drawing upon human rights and UBN. The United Nations Development Programme (UNDP) global acute poverty measure in collaboration with the Oxford Poverty and Human Development Initiative (OPHI) that recovers some key dimensions of the UBN approach and aggregates the indicators using the AF methodology (Alkire & Santos, 2010; UNDP, 2014). The most recent regional example is the EUROSTAT deprivation index that draws on relative deprivation and the consensual method (Guio et al., 2017, 2016).

Only two official measure have been put to statistical scrutiny: EUROSTAT and CONEVAL. The EUROSTAT measure was produced by Guio et al. (2012) and revalidated by Guio et al. (2017). It is based on the consensual approach and relative deprivation theory and has been fully statistically validated. The reason why there are several arrows pointing at the CONEVAL measure (the official Mexican measure) is that this is arguably the first official multidimensional measure and it had a desirable process for its production. After a critical change of the Social Development Law in 2004, Mexico was obliged to measure poverty from a multidimensional perspective and in accordance with the constitution. This resulted in an international consultation process. Mora (2010) compiled the different contributions of the authors and CONEVAL (2011b)

summarises the contributions of each author. The resulting measure is a hybrid measure that benefit from an international collaboration and it has been the first one to be fully statistically validated before its production. Townsend (1979) and the consensual approach were used to assess the thresholds for the nominal variables and provide a view on the socially perceived needs of the Mexican population with focus on social rights. Then Gordon (2010) conducted a preliminary evaluation of the indicators of the measure and suggested an approach to identify the poor and the not poor. Then the AF method was used to estimate the depth and intensity measures for income and the deprivation score.

Both EUROSTAT and CONEVAL have succeeded in producing measures that are not only theoretically sound but are also empirically validated. This is, nonetheless, an exception to the current practices in poverty research. The remainder of the book thus focuses on how to avoid producing magnifying-noise scales.

Chapter 2

Poverty and measurement theory: A statistical framework

Abstract

This chapter outlines a statistical framework to tackle some of the challenges that involved in multidimensional poverty measurement. Measurement theory is posed as overarching framework for the assessment of some of the key assumptions made in the production of multidimensional poverty scales. The concepts of reliability, validity, measurement invariance and scaling are defined and put into the context of poverty research. Working examples of this concepts are then presented in the following chapters of the book.

2.1 Work flow in poverty measurement: A falsifiable framework

Chapter 1 outlined some of the tasks involved in the development of multidimensional poverty indices and highlighted the assumptions underlying each stage: dimensions, indicators and thresholds or cut-point selection; weighting of dimensions and indicators; aggregation or production of a score and setting of a poverty line Alkire (2007), Thorbecke (2007) and Gordon & Nandy (2012).

Figure 2.1 illustrates the strategy often followed by researchers to produce a poverty index. Data often precedes the measurement (data ara given). Therefore, researchers have little influence upon the data collection process and thus they are constrained and have to adapt the existent information on deprivation to their poverty definition -or in the worst cases, they have to adapt the definition to the data-. A series of assumptions are raised with regard the number and type of dimensions, the indicators (including cut offs to identify deprivation) and weighting (see Chapter 1. Ideally, these assumptions should be assessed using a framework but in practice researchers avoid this stage or they conduct a series of ad hoc sensitivity analyses. The limitation is that these kind of analyses have no hypothesis and researchers are more likely to confirm their beliefs due to the lack of a clear testable strategy. Confirmation biases, therefore, are more likely to remain and basically researchers jump from their theoretical measure to the aggregation procedure.

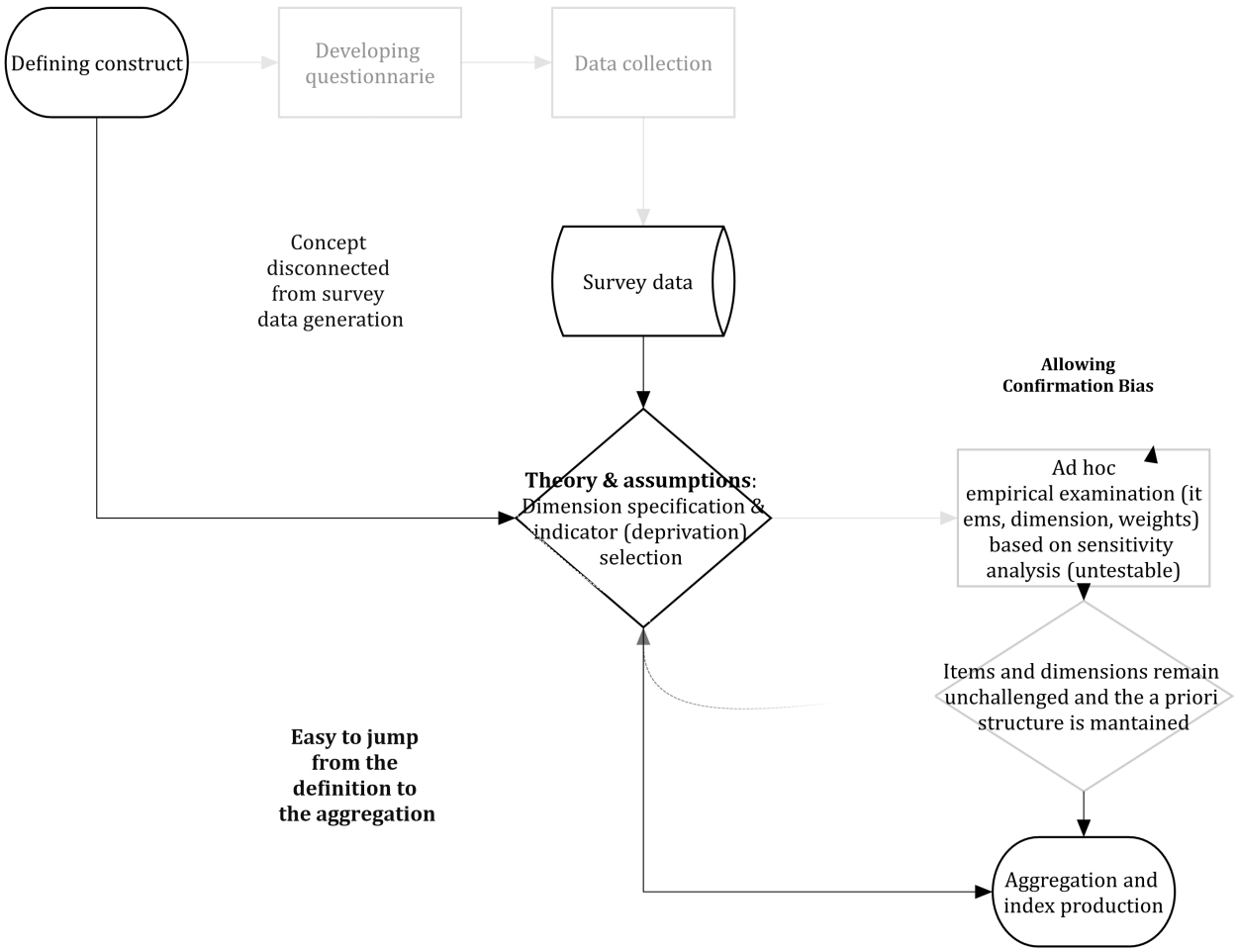

Figure 2.1: The work flow researchers often implement in multidimensional poverty measurement.

Scientific measurement aims to incorporate a framework to falsify researcher's assumptions. Figure 2.2 shows the strategy that researchers could employ to reduce confirmation biases and measurement error. A theory of the concept of poverty should guide the development of a survey questionnaire and data collection. Researchers then could further especify the structure of their poverty measure based on the theory-driven data generation process. Once the theoretical measure is defined, assumptions need to be identified and make explicit so that can be falsified using a sound statistical framework. The results can be used as an input to redefine the measure in an iterative process. Once the measure is proven to be robust (using a definition of what a good measure is derived from a measurement framework), researchers could move onto aggregation and the identification of the poor and not poor groups.

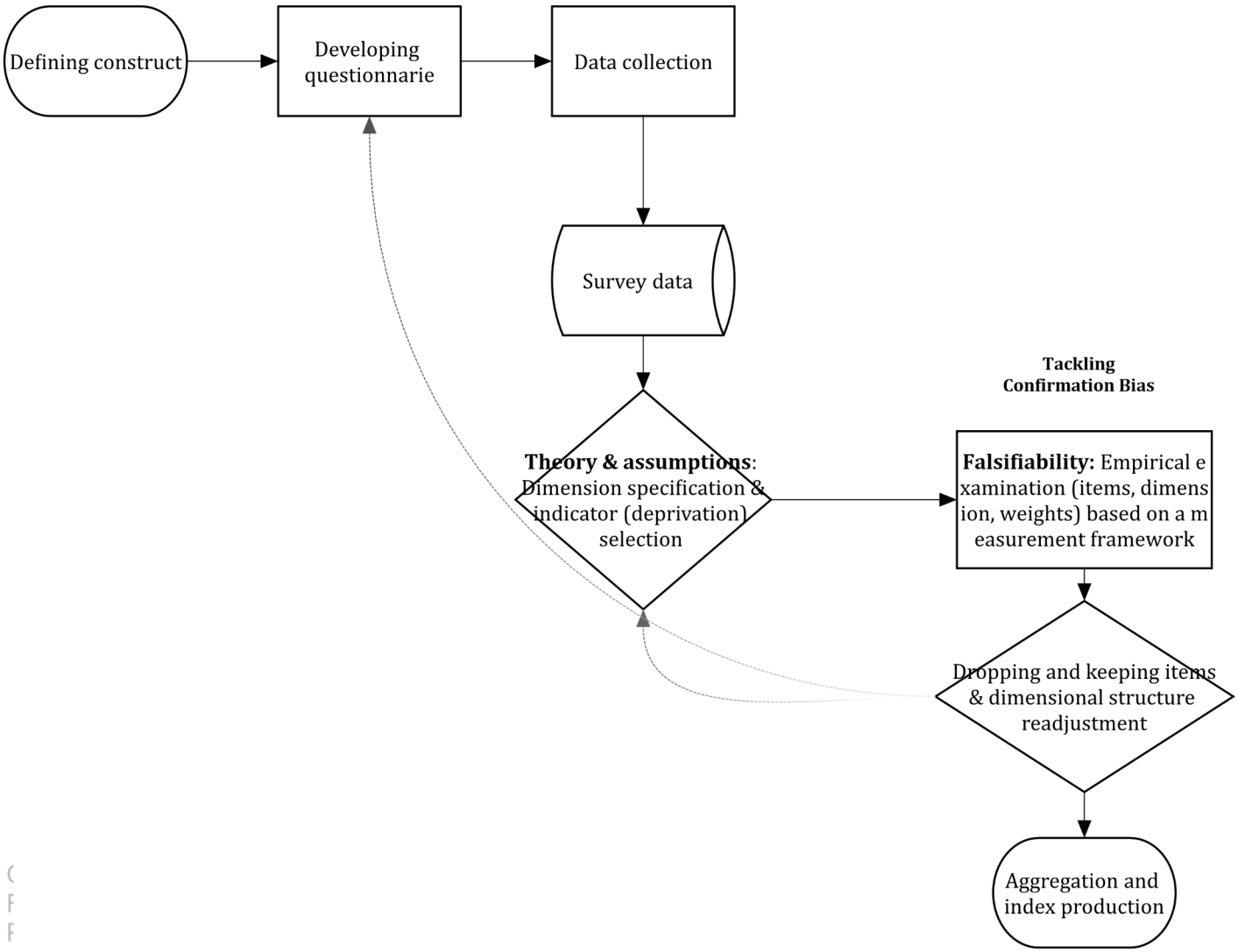

Figure 2.2: Ideal work flow in multidimensional poverty measurement.

One possible and useful way to advance in the empirical assessment of poverty indices consist in detecting the assumptions involved in the production of an index and translate them into a falsifiable framework. That is, raising questions about all researcher's assumptions and propose a method to answer whether researcher's ideas hold given the data.

The challenges involved in poverty research can be put in terms of sequential stages following list of problematic or contested issues in poverty measurement. All these issues are part of an effort to approximate the extent and identification of poverty and can be therefore written in terms of a statistical model: a simplified and imperfect description based on observed data. As McCullagh (2002) argues the idea of a statistical model conveys the recognition that the characterization of something is just an approximation. As such, a model to measure poverty is one possible option that can be assessed and improved. Before saying how a poverty measurement model can be examined and improved, is important to present an organised falsifiable framework of the key assumptions in poverty measurement.

2.2 Identification of the sampling space

One of the often overlooked features in poverty measurement is that fact that researchers never work with the full set of information. The main reason is that such set is unknown or unavailable. Given a definition of poverty, there are different dimensions, indicators and parameters to produce a working model to approximate poverty. But all these option belong to a space with all possible options. This is a major aspect of poverty measurement and highlights the fact that there is an underlying assumption (based on theory or data) about what subsets would work better to measure poverty.

1. Sampling space of all possible dimensions $\mathscr{J}$
2. Sampling space of all possible variables $\mathscr{X}$
3. Sampling space of all possible parameters (e.g. weights) Θ

This is just a statistical model throughout which poverty is identified and can compactly be written as:

$$\mathscr{F} = \{\mathscr{X}, F_\theta : \theta \in \Theta\} \tag{2.1}$$

where the variables $x_1, ..., x_n$ follow a certain distribution F_θ, which is indexed by a parameter θ defined in the parameter space Θ. $\mathscr{F}$ is a family of all probability distributions on $\mathscr{X}$, which is just the set of all possible observed data.

2.3 Selection of dimensions and indicators

Poverty measurement therefore involves sampling from different spaces (array of options). For example, they put forward some dimensions, variables and weights. This is often the first challenge in poverty measurement where researchers select some dimensions j from the sample space $\mathscr{J}$ and some x_{ij} from the sample space $\mathscr{X}$, so that if the scale has 30 indicators $\mathscr{X} = 1, 2, ..., 30$ where $x_{ij} = 1$ when deprived and $x_{ij} = 0$ (and $z > 0$ for binary variables). For nominal variables, there is a space of possible cut offs $\mathscr{Z}$ too.

Selecting dimensions, indicators and thresholds involves making the assumption that the sampling from the different spaces is the best possible one, i.e. the one that leads to a good poverty measure How does a researcher know whether its selection(sampling) is not wrong? This can be broken down into several questions:

- Is the subset of dimenions j from $\mathscr{J}$ an *adequate* characterization of poverty?
- Is the subset of indicators given a cut off $(X; z)$ from $\mathscr{X}$ an *adequate* characterization of the dimension j and poverty?

The word *adequate* is a loose term as this point as we need a theory or standard to define it. This is covered in the next section but at this point, the focus is on the into translating the challenges in poverty research into assumptions to build a statistical model (that later can be testable in some way).

2.4 Aggregation and weighting

The first stage is focuses in indicator and dimension selection and as it will be discussed below, this is where a falsifiable framework is more useful but also absent in poverty research. One way to illustrate it is with the AF method. The AF method will work fine as long as the components of the formula work fine but there is nothing within it that will ensure that this will be the case. Therefore, once the indicators have been selected, in a second state, researchers aggregate the variables using a linear model selecting some weights $w \in \mathscr{W}$, where $w_{ij} = 1$ for non-differentially weighted measures and $w_{ij} \neq 1$.

- Does the weighting scheme lead to the same ranking of the population?

2.4.1 Splitting the population into meaningful groups

Once the model of poverty has been completed $\mathscr{F}$ a score is produced for each person in the sample. That means that a threshold should be proposed to identify the poor population.

2.5 Measurement theory as an statistical framework

2.6 Poverty and error in measurement

Poverty, as happens with many other constructs in social sciences, is an idea. It emerges from the theoretical presumption that within a population there is a group of people whose livings standard are below of what a society at a given point in time consider essential and customary to have a decent life. Poverty, in theory, impedes full participation in society, enhances the risk of die younger, interacts with many social risks, etc. But poverty is difficult to pinpoint because it is not directly observable and cannot be described with accuracy by a single variable. Instead, poverty research must rely on imperfect multivariate data to rank the population according to their living standards. To make things more difficult (see previous section), there are many ways in which poverty can be captured or described as researchers rely on different sets of outcomes (different samples from the same space) to characterise one's living standards (diverse deprivation/needs/achievements). Poverty thus is a latent variable in that several outcomes are utilized to imperfectly describe it[1].

The problem of capturing poverty as a construct is just one of many cases in social sciences where researches face the difficulty of finding a theoretically constructed (unobservable) group (e.g. the depressed, high class, high-achievers in education, happy). How then can we detect a group of people belonging to a construct? Spearman (1904) put forward a capital idea: two or more outcomes is an indication that two things may have the same underlying cause. This perfectly fits the theory that deprivations are an observed outcome of poverty. A core assumption in poverty measurement is that a set ($X \in \mathscr{X}$) of indicators constitutes the series of relevant manifestations of the idea of poverty. This treatment of poverty as a construct is powerful in that is a theoretical underpinning and rationale for finding and ranking individual differences in living standards (below which one is poor).

Deprivation indicators are intercorrelated because they share a common cause: poverty. Conceptually, this means that if the effect of poverty is eliminated the inter-correlation of outcome variables would be zero. Measurement theory postulates that the problem of the relevant set of outcome variables to measure poverty can be tackled from the perspective of the *common factor model* Thurstone (1947). Such a model formally postulates that the observed outcome is a function of one or more common factors and one unique factor. Each deprivation should vary due to two main sources (1) common variance and (2) unique variance. The first type of variance is the one accounted by for the latent factor- the variance shared with the other outcome measures. The second is the variance accounted by for other factors and by random error (unreliability, measurement error).

2.7 Measurement model for poverty

The statistical model proposed in modern measurement theory would be a common factor model where each observed variable (x_{ij}) is a product of a latent dimension (η_j) and the higher order factor (ζ_h) overall poverty. One of the interesting aspects of this formulation is that as not everything is due to the latent variable, error theory of a measurement model includes the method effect.

$$x_{ij} = \lambda_{ij}\eta_j + \varepsilon_i j \tag{2.2}$$

$$\eta_j = \gamma_j \zeta + \xi \tag{2.3}$$

[1]There is another framework that incorporates the idea of measurement error. Fuzzy sets theory uses the term vagueness to describe the intrinsic imperfection in poverty measurement (Martinetti, 2006)

The λ_{ij} and γ_j are known as factor loadings and they capture the relationship between the latent variables and the outcome measures, and between the dimensions and the overall latent variable. Of course, it is possible to have more ζ's but in poverty research the presumption seems to be that dimensions are nested into one overall latent construct which is poverty (some especial cases will be shown in Section XX where the higher-order model is a third-order factor):

Poverty → Dimensions → Sub-dimensions → outcome measures.

This model specification is one way to statistically capture the way in which multidimensional poverty measurement is conducted in the contemporary literature. Researchers propose a series of dimensions and classify their proposed indicators accordingly. The crucial aspect is that this is just an idea -informed from some theory or data- about how poverty can be better captured under a multidimensional definition. That is, the proposal is just an invalidated model that demands scrutiny. When put in terms of equation (2.2) and equation (2.3) researchers move from theoretical speculation toward empirical falsification.

2.8 Blueprints and poverty measurement models

Models are just blueprints that summarise an explanation of how things can be constructed. Putting theoretical proposals to measure poverty in terms of a diagram helps a lot to visualise and contrast the diverse proposals to measure poverty. The advantage of measurement theory is that everything can be put in terms of a blueprint. Multidimensional models can be easily presented as a diagram, which is just a graphical representation of a model. Another way to think about this, is to see the different blueprints in poverty measurement are just the different ways in which researcher sample from $\mathscr{J}$ and $\mathscr{X}$. They lead to diverse structures, i.e. models.

Equation (2.2) simply tells a model in which one latent construct produces the observed (x_{ij}) outcomes. This model is say to be unidimensional as there is only one factor η_j causing the observed indicators. One could think of this model as the null model in poverty measurement where the indicators cannot be clearly clustered into dimensions. This does not mean that the indicators do not measure different aspects of poverty. It might be that the indicators indeed different aspects but that there are no clusters of indicators. In practice, there are few theoretical models proposing such thing. Empirically, however, it might be often a case given that in poverty research data collection rarely follows a theoretical proposal (see *add chapter reference*). Figure 2.3 translates (2.2) into a plot.

Figure 2.4 displays the higher order factor model using the three dimensions proposed by A Guio et al. (2009). This structure is roughly what poverty researchers have in mind when thinking about poverty in multidimensional terms. That is, that the indicators can be grouped into some dimensions. In this example, there are 9 outcome variables classified into three dimensions (η_j) (Durables, Housing and Economic strain). Then the loadings (λ_{ij}) denote the relationship of each outcome with the dimensions in question. Then the arrows from the high-order factor (overall poverty) to each dimension are the factor loadings (γ_j) that capture the relationship between overall poverty and each dimension. Both (λ_{ij}) and (γ_j) are parameters of the model and denote the strength of the association between the latent variable and the outcome variables (See **??** for an explanation of how this relates to the topic of reliability and differential weights). The diagram of @(fig:cfaguio) model is an example of a higher-order factor: A three-dimension model with a higher-order factor. The dimensions are durables, housing and economic strain. In this example, there are only three indicators for each dimension. In this model, none of the indicators loads into more than one dimension. Poverty measures often assume that a given indicator is an exclusive outcome of a certain dimension.

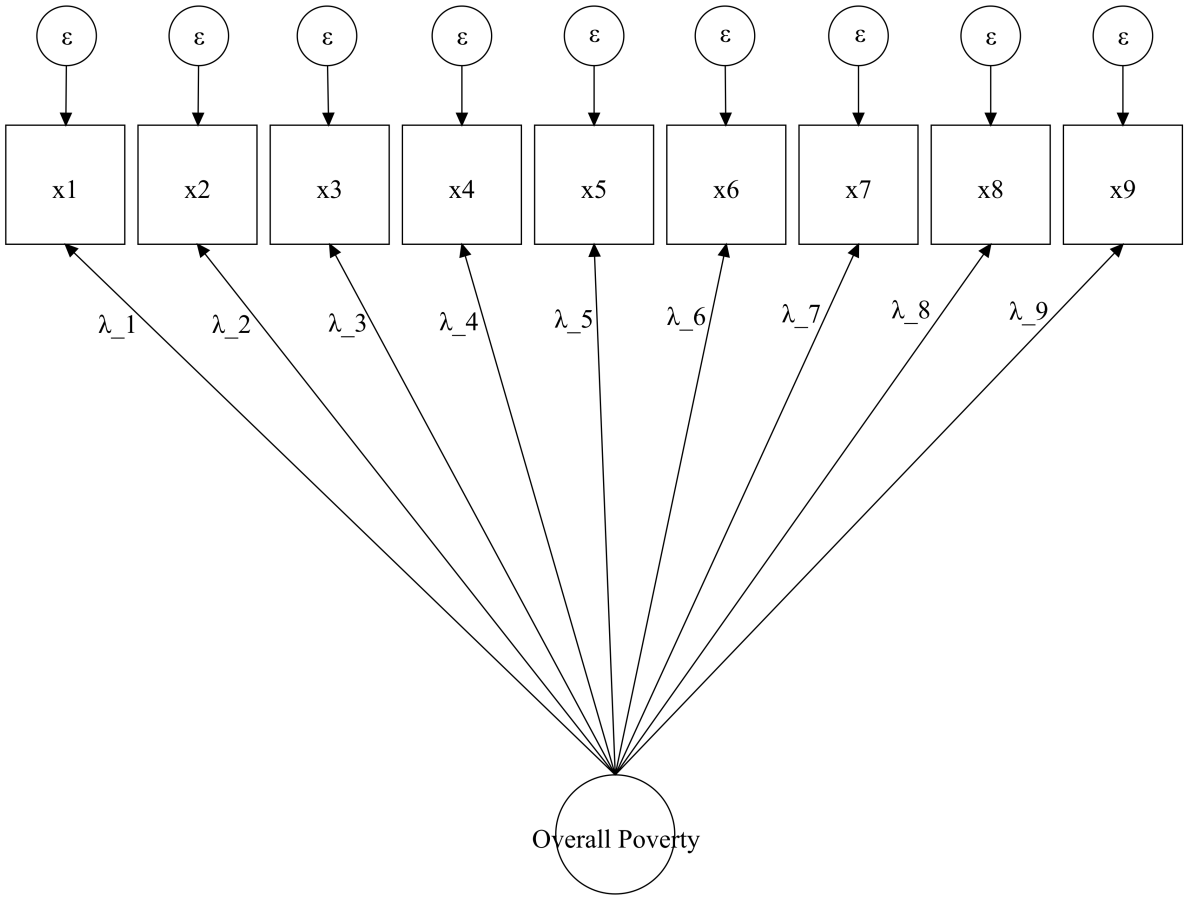

Figure 2.3: This is a visual representation of a null unidimensional model.

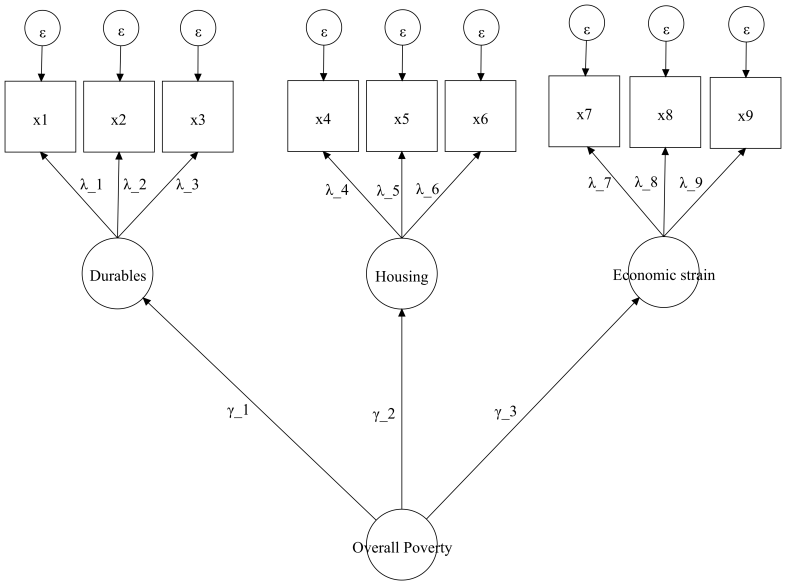

Figure 2.4: This is a visual representation of @Guio2009's model. Second-order factor.

Another way to think about the dimensions of poverty comes from Alkire & Santos (2010), which is used to compute OPHI-UNDP's Multidimensional Poverty Index (MPI). The structure proposes a similar structure to Guio et al. (2009)'s model: second-order structure. In this case, poverty is thought to have three substantive dimensions: Education, health and iving standards. These are different dimensions from those porposed by Guio et al. (2009) or by Townsend (1979) in figure 2.5. The diagram only specifies, at this point, the structure of the measure as the MPI does not have three indicators for each dimension but it helps to see the model the authors put forward to measure global acute poverty.

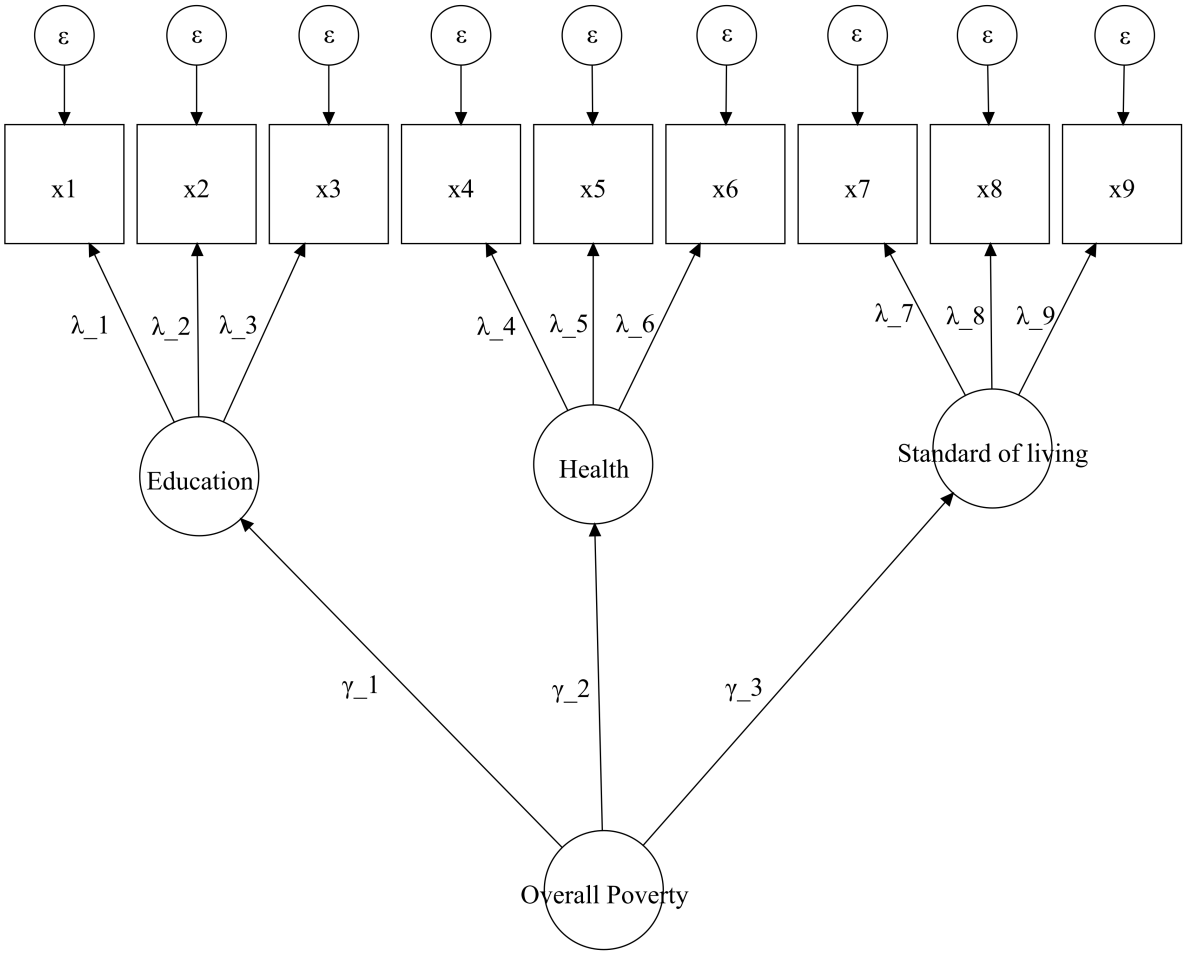

Figure 2.5: This is a visual representation of @Alkire2010's model. Second-order factor

Townsend (1979)'s model is one of the first multidimensional models in the world. His proposal has more nested dimensions compared with figures 2.4 and 2.5. In this case, indicators are nested into eleven dimensions, which in turn can be grouped into two more dimensions. The resulting figure (2.6 is a third-order factor structure. Guio et al. (2017) propose a reduced version of this model which does not considers the 11 dimensions and indicators are classified according to material and social deprivation. The loadings of the indicators are ommited. κ_1 and κ_2 are the loadings of the higher order factor. These could be specified using (2.3) as reference- now ζ has k=2 factors.

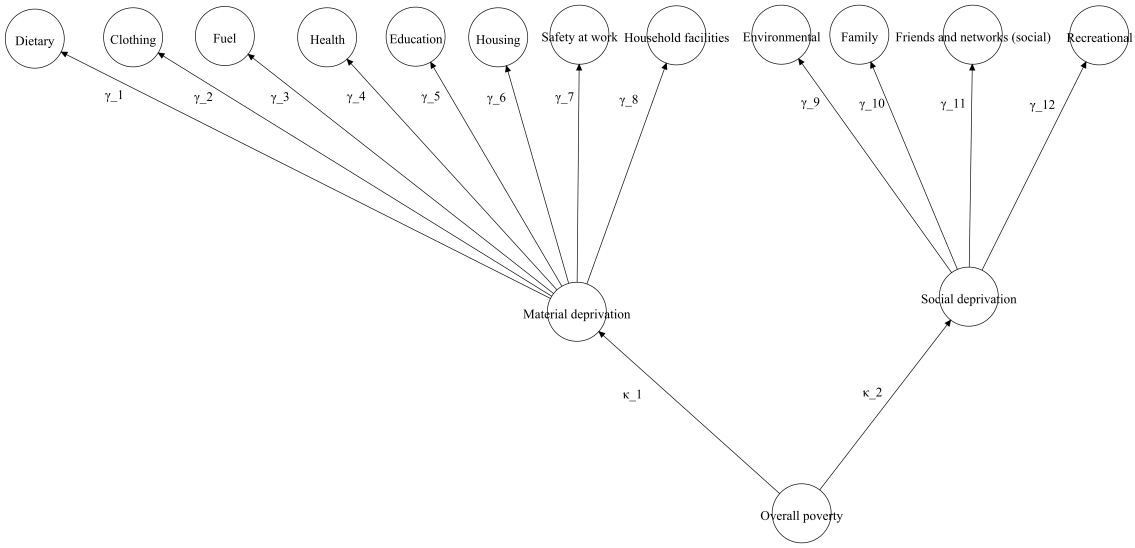

Figure 2.6: This is a visual representation of @Townsend1979's model. Third-order factor.

Another way to think about this, is to see the different blueprints in poverty measurement are just the different ways in which researcher sample from $\mathscr{J}$ and $\mathscr{X}$. They lead to diverse structures, i.e. models. This reflection brings us back again to the question: how do we know whether the samples we take from $\mathscr{J}$ and $\mathscr{X}$ are an adequate representation of poverty in a given society at a given point in time?

2.9 Measurement theory and principles

One central debate in poverty measurement is about the dimensions of poverty. The discussion revolves around questions such as: How many dimensions are? What are the contents (indicators) of these dimensions? How dimensions are associated and differentiated? Does these dimensions have equal importance? (see **??**). The previous section translated these questions into concrete challenges and advanced a general framework of measurement that explicitly acknowledges that all measurement practices involve different kinds of error. However, the framework is incomplete. It requires some governing principles that effectively put the work flow in poverty measurement in terms of a cogent falsifiable framework. One crucial question raised in section~**??** are:

- Is the subset of dimenions j from $\mathscr{J}$ an adequate characterization of poverty?
- Is the subset of indicators given a cut off $(X; z)$ from $\mathscr{X}$ an *adequate* characterization of the dimension j and poverty?
- Does the weighting scheme lead to the same ranking of the population?

2.9.1 Origins of measurement theory

Measurement consists in assigning a series of numbers to individuals in such a way that they represent quantities of attributes (Nunnally & Bernstein, 1994). *Measurement theory* is a framework that postulates that a series of outcome measures are manifestations of a latent trait in each individual. That is, a framework to turn a series of deprivation indicators into numbers so that they express the unobserved level of poverty of an individual. Hence, measurement theory aspires to provide a series of rules to distinguish signal from noise so that our observed measures approximate the latent trait in question.

The origins of measurement theory can be traced back to classical test theory (Lord, 1952; Novick, 1966). This theory postulated that a true score is a linear combination of an observed score and error. This idea has been

taken forward by the latent variable approach through a series of breakthroughs in theory, conceptualisation via latent constructs and computation (Cudeck & MacCallum, 2012; Rusch, Lowry, Mair, & Treiblmaier, 2017). The consolidation of this framework required parallel and temporarily disconnected contributions in factor analysis and item response theory (IRT). After Spearman (1904)'s seminal one factor model and Thurstone (1947) multiple factor contribution, as series of works in the 1960s proposed formulating factor analysis not in terms of a correlation matrix but in terms of a model Lazardfeld & Henry (1968);Lawley & Maxwell (1971). Later and independently, it seems, from factor analysis, in educational testing and psychometrics, a series of works from Stocking & Lord (1983) and Bock & Aitkin (1981) proposed item response theory (IRT) which proposes that indicators are measures of observed manifestations of an underlying trait (Reise, 2014). The birth of IRT was almost contemporary to Jöreskog (1970);Joreskog, Sorbom, & Magidson (1979)'s contributions to confirmatory factor analysis and structural equation models. These factor models could be seen now as general case of IRT for categorical variables (Muthén, 1984).

Modern measurement theory is a more unified framework that postulates that outcome measures are manifestations of a latent trait, that such manifestations could be clustered together into sub-dimensions, that measurement will always have error and that all these aspects should have an empirical counterpart to that can be tested (Cudeck & MacCallum, 2012).

Measurement theory since the seminal work of Spearman (1904) has continuously developed a cogent framework that aims to produce measures that: 1) consistently offer the same ranking of a population and 2) a ranking actually 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*. Furthermore, these two have important implications in terms of weighting, comparability and identification of (unobservable) population groups.

Chapter 3

Reliability in poverty measurement

Abstract

This chapter introduces the theory and concept of reliability. An intuitive explanation is provided at the beginning of the chapter to underline the implications of reliability for measurement. Then, a formal introduction to the theory of reliability is provided. Reliability can be estimated using different approaches, the chapter discusses its limitations. The second main section of the chapter illustrates how reliability works and how can it be estimated using **R** and **Mplus** by using simulated data. Then a real-data example is used to show some of the typical problems involved in the examination of reliability.

3.1 Intuition to the concept of reliability

In what sense the concept of reliability relates to the idea of having a measure we can trust? Poverty analysts and policymakers require indices they believe in to focus on more important issues like developing and studying poverty eradication strategies. 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 build upon consistent and meaningful estimates. For example, imagine a case in which we could conduct two surveys to the same population. Ideally, we expect our classification of the poor population to remain unchanged from t_1 to t_2. This is telling us that our index is stable across samples. A noisy index, in contrast, would lead to unstable orderings and it is impossible to distinguish a signal (the thing we are interested in) from noise (unnecessary and confusing variability).

Consistency, however, is not simply having the same response patterns *ceteris paribus* across two samples but also by having systematic population orderings. 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) tells us to be suspicious about such kind of behaviour. 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" See Chapter 1). The consequence would be 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. Therefore, even with a very similar response pattern, our scale will be rather unstable to be trusted. Of course, if we know that the folding bicycle item is an unreasonable measure of poverty we would have drop it before the empirical analysis. However, in poverty measurement there are variables or thresholds of these variables that are quite contested.

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 dropped this indicator from our analysis we would lose valuable information. Because we will be missing good variables (either because are not available or we just miss it from theory) we would like a measure whose population ordering is not that sensible to

information losses. That, indeed, is a measure we can trust in the sense that it will lead to consistent results. 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.

3.2 Reliability theory

Reliability is a key concept in measurement theory and can be simple defined as the homogeneity of an index (Revelle & Zinbarg, 2009). An homogeneous index is a scale whose outcome indicators are manifestations of the same trait. In the literature several authors refer to reliability as internal consistency of an index because this a consequence of homogeneity. In the example above having an indicator that is not a good measure of poverty means that the index is heterogeneous 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.

The theory of reliability is inextricably connected with the evolution of measurement theory. The theory of reliability can be traced back to classical test theory (CTT) but it has been under continuous development by more recent breakthroughs in latent variable modelling. Reliability is rooted in the acknowledgement that all measures have an unknown mixture of signal and noise (error). For Spearman (1904) there should be a *true* score- that is just the combination of an observed score and error. As in classical or frequentists statistics, it is *true* in the sense of the expected score across many replications of the same experiment. Being θ the true score, in CTT reliability is expressed as:

$$x_i = \theta_i + \varepsilon_i \tag{3.1}$$

Equation (3.1) can be put in terms of variance decomposition. The variance of the observed score σ_x^2 is thus equal to the variance of the true score plus the variance of the error. The discrepancy between the true score and the observed score is an estimate of reliability:

$$\rho = \frac{\sigma_\theta^2}{\sigma_x^2 + \sigma_e^2} \tag{3.2}$$

where ρ is the total reliability and σ_i^2 is the subject's variability and σ_e^2 is the measurement error. Because this is a simple proportion, the reliability estimate will be (almost) always between 0 and 1[1].

The classical definition of reliability has been translated and adopted by the latent variable approach. This approach is not at all concerned with the *true* score but with the extent to which a measure reflects the construct. Here the factor loadings λ_i's are key in that they reflect the association between an outcome and the latent construct. Therefore, latent variable approach naturally accommodates the question about how good are the manifest variables. Furthermore, it can estimate both σ_x^2 and σ_e^2. Reliability can be expressed as:

$$\rho_{x_i\theta} = \frac{\lambda_i^2}{\sigma_x^2} \tag{3.3}$$

3.3 Statistical measures of reliability

The are different ways to estimate the reliability of a scale, each one with its advantages and disadvantages. The most widely use estimate of reliability is α or λ_3 (do not mistake with factor loadings) (Cronbach, 1951; Guttman, 1945). This estimate comes from CTT and draws upon Spearman (1904) approach to estimate the variance based on parallel tests:

[1] If the scale is badly constructed reliability could be negative using some statistics like α

$$\alpha = \lambda_3 = \frac{\sigma_x^2 - \sum \sigma_{xi}^2}{\sigma_x^2} \frac{n}{n-1} \tag{3.4}$$

Cronbach's α is, nonetheless, not a good estimate of reliability (Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005). 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, the outcome measure have equal error variances. These two assumption 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. Table~**??** summarises the relation among the different reliability statistics by dimensionality.

Given that α is based upon untenable assumptions, there have been several proposals to estimate reliability under more general conditions. Revelle (1979) proposes the statistic β. This coefficient considers the worse split in different halves, i.e. it minimizes the average covariance by taking into account the lowest inter-item correlation ($\bar{\sigma_{ij}}$). It is thus a measure of the lowest possible reliability and therefore it will always be lower or equal to α. It is estimated as follows:

$$\beta = \frac{k^2 \bar{\sigma_{ij}}}{\sigma_x^2} \tag{3.5}$$

McDonald (1999) put forward two alternate measures of reliability: ω and ω_h. The first statistic is also know as the measure that maximizes the estimation of reliability, i.e. the lowest upper bound (Zinbarg et al., 2005). Equation (3.6) shows the formula of ω. This equation is a proportion of the variance of the latent variable that is accounted by the outcome measures.

$$\omega = \frac{\sum_{j=1}^{k} \left(\sum_{i=1}^{p} \lambda_{ij} \right)^2}{\sum_{j=1}^{k} \left(\sum_{i=1}^{p} \lambda_{ij} \right)^2 + \sum_{i=1}^{p} e_i} \tag{3.6}$$

Equation (3.7) shows the formula to estimate ω_h which is also a proportion but in this case is the variance accounted by the higher order factor. Therefore, this is a more appropiate measure when having multidimensional scales.

$$\omega_h = \frac{\left(\sum_{i=1}^{p} \lambda_{ij} \right)^2}{\sum_{j=1}^{k} \left(\sum_{i=1}^{p} \lambda_{ij} \right)^2 + \sum_{i=1}^{p} e_i} \tag{3.7}$$

These different reliability statistics beg the following question: Which one should be used? There are two complementary ways to answer this question. First, these reliability statistics are based on a series of assumptions and thus its usage depends on the extent to which each one is adequate given the data and the research question.

α is a very specific case whose assumptions will be rarely meet in practice. The recommendation is to avoid using α and focus on general cases such as ω and ω_h. ω will work in almost any situation but when the measures are multidimensional. This does not mean that it would be incorrect to use it. In multidimensional settings, ω_h is just more adequate because it will tell the amount of variance accounted by for the higher order factor.

Zinbarg et al. (2005) ran a Monte Carlo study to assess how does the different reliability statistics compare one another. They found the following (See Table 3.1):

Table 3.1: Summary of the relations among β, α, ω and reliability depending on index dimensionality. Taken from (Zinbarg et al., 2005, p. 128)

Dimensionality	Expected behaviour
Multidimensional	$\beta < \alpha < \omega \leq \rho$
	$\omega_h < \omega \leq \rho$
Unidimensional	$\beta < \alpha < \omega_h = \omega \leq \rho$

The second way to answer the question has to do with the conclusions one could make from the estimation of these measures. If the assumptions are violated our conclusions would be very likely incorrect and misleading. Assuming the correct statistic is selected, the question is: How low is too low to be unacceptable?

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. The summary of the findings of Nájera (2018) are shown in Table 3.5.4. The simulation considered three possible dimensional structures: unidimensional, weak and strong multidimensional measures. Weak multidimensionality was defined as the case where the dimensions have relatively low loadings to the higher-order factor.

Table 3.2: Summary of the relations among β, α, ω and entropy depending on index dimensionality. Summarised from Nájera (2018). In this case, the unidimensional model seem to meet τ equivalence, i.e. equal loadings.

Reliabiity statistic	Leads to	lassification error (%)	Entropy value
$\alpha > .8$	$\approx$	$< 5\%$	$> .8$
$\omega > .8$	$\approx$	$< 5\%$	$> .8$
$\omega > .85$	$\approx$	$< 5\%$	$> .8$
$\omega_h > .65$	$\approx$	$< 5\%$	$> .8$
$\omega > .85$	$\approx$	$< 5\%$	$> .8$
$\omega_h > .70$	$\approx$	$< 5\%$	$> .8$

3.4 Item-level reliability and weighting

Classical test theory was concerned with overall reliability. Item response theory (IRT) move 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. This kind of model has been used by Guio et al. (2016) and Guio et al. (2017) for example.

$$P_i\theta = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{3.8}$$

Equation (3.8), translated to poverty measurement, 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) show how this models relates to a unidimensional factor model, equations 21 and 22. In a factor model (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.

$$a_i = \lambda_i \sqrt{\psi} \tag{3.9}$$

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. However, this is no longer the case as it is possible to estimate multidimensional IRT model (Reckase, 2009). However, Gibbons, Immekus, Bock, & Gibbons (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.

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 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.

One of the most contested issues in poverty measurement revolves around weighting (Decancq & Lugo, 2013). Measurement theory proposes that reliability lead to a self-weighting measure in that it guarantees good population classification (Streiner et al., 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.

One of the key axioms in poverty research is the monotonicity axiom. It states that poverty *ceteris paribus* should decrease after an improvement in one's achievements (Alkire et al., 2015; Sen, 1976). Measurement theory states something very similar in that low loadings reflect the fact that changes in the latent variable do not lead to changes in observed deprivation. Nájera (n.d.) ran a Monte Carlo experiment the particularities of this behaviour. He finds that item-level unreliability leads to a violation of the monotonicity axiom. His conclusion is that indicators that have weak discrimination $\lambda_i j < .4$ (standardised loadings) violate weak monotonicity and in some circumstances could violate strong monotonicity. Therefore, such indicators are more noise than signal to poverty measures.

3.5 Estimation of Reliability

3.5.1 Overall reliability

To introduce the idea of reliability we will use the data set "Rel_MD_data_1_1.dat". This is simulated data of a higher-order multidimensional measure of poverty ($n = 5000$). The measure has nine indiactors in total distributed evenly in three dimensions.

```
library(plyr)
Rel_MD_1<-read.table("Rel_MD_data_1_1.dat")
```

```
Rel_MD_1$ds<-rowSums(Rel_MD_1[,c(1:9)])
colnames(Rel_MD_1)<-c("x1","x2","x3","x4","x5","x6",
                      "x7","x8","x9","x10","x11",
                      "resources","educ_yr","occupation",
                      "class","hh_members","ds")
Rel_MD_1[1:10,1:11]
```

```
##    x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11
## 1   1  1  1  1  0  0  0  0  0   0   0
## 2   0  0  0  0  0  0  0  0  0   0   0
## 3   0  0  0  1  0  0  0  0  0   0   0
## 4   1  1  0  0  0  0  1  0  0   0   0
## 5   1  0  0  0  0  0  0  0  0   1   1
## 6   1  0  0  0  0  0  0  0  0   0   0
## 7   0  0  0  1  0  1  0  0  0   0   0
## 8   0  0  0  1  0  0  0  0  0   1   1
## 9   1  0  0  1  1  1  1  1  1   0   0
## 10  0  0  0  0  0  0  0  0  0   1   0
```

We do not know yet if our selected deprivation indicators lead to a reliable score. However, we can inspect its distribution by plotting it (Figure 3.1) as follows:

```
require(ggplot2)
ggplot(Rel_MD_1, aes(ds)) +
    geom_histogram() + theme_bw() + labs(x = "Deprivation score") +
  scale_x_continuous(breaks = seq(0, 9, by = 1))
```

Now we can check the proportion of people deprived of each indicator as follows:

```
dep_prop<-unlist(lapply(Rel_MD_1, function(x) mean(x)))
dep_prop<-round(dep_prop[1:9]*100,0)
dep_prop
```

```
## x1 x2 x3 x4 x5 x6 x7 x8 x9
## 50 29 16 49 29 16 45 26 16
```

3.5.2 Exploratory (non-model based) estimation of overall reliability

ow that we have familiarised with the data ourselves we can proceed to check the reliability of this scale. Reliability concerns with the homogeneity of a scale and its capacity to produce consistent rankings of a population. We will start by estimating the overall reliability of our scale using the `psych` package (Revelle, 2014). This is a comprehensive R-package to estimate different reliability statistics (α, β, ω and ω_h) under different changing conditions. The `psych` package can be used for exploratory and confirmatory settings for both unidimensional and multidimensional measures. This book focuses on confirmatory measurement models and to introduce the estimation of overall reliability we will rely on the simplest way to estimate the homogeneity of a scale using the simulated data set. This will be further developed and the next section shows how `psych` interacts with another R-package `lavaan` to estimate ω and ω_h from a confirmatory factor model (Rosseel, 2012).

The `pysch` package permits estimating α and ω using the same function (`omega`). The package has several options but we know that there are three dimensions and one higher order factor and these values match the defaults of the `omega` function. It is important to bear in mind that in this simple case the value of ω is approximated with an Exploratory Factor Analysis (EFA). Below is shown how to do it with a confirmatory model.

After applying the `omega()` to our nine indicators, there will be different objects that store information with the results of the analysis. We will focus on the overall estimate of α and ω as here we are interested in

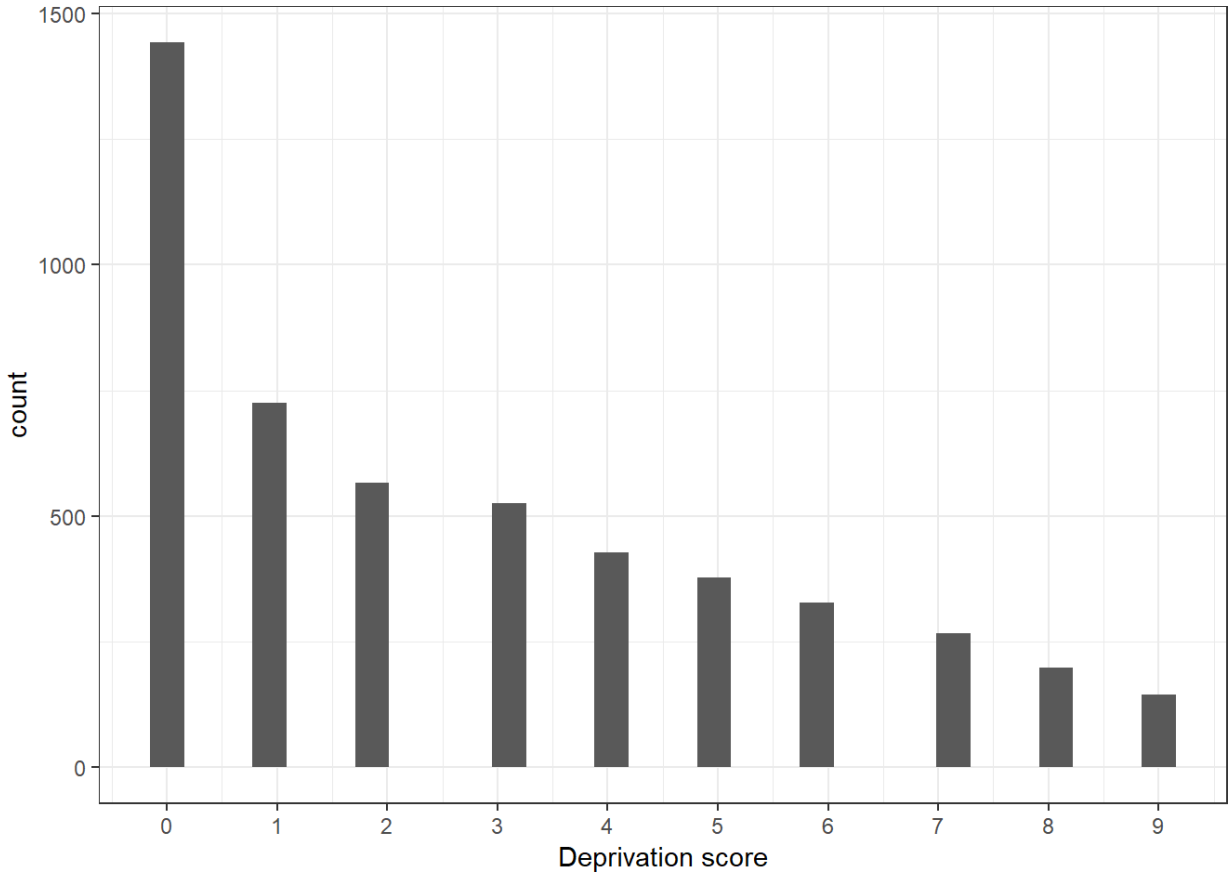

Figure 3.1: This is the histogram of the deprivation score. It shows the number of people by the equally weighted deprivation count.

knowing the homogeneity of our scale. We can appreciate below that both values are high ($\geq .8$) (See *ref* for an explanation) and suggest that the scale is highly reliable. In this case, $\alpha < \omega$ indicating that this scale violates τ equivalence (equality of loadings).

```
# install.packages("psych")
require(psych)
omega_exp1<-omega(Rel_MD_1[,c(3:9)])
rel_uni_exp<-data.frame(omega_exp1=omega_exp1$omega.tot,
                        alpha=omega_exp1$alpha)
rel_uni_exp
```

```
##   omega_exp1     alpha
## 1  0.8599319 0.8127129
```

Both α and ω are easily estimated with the `psych` package. However, the previous example was pretty straightforward in that all the indicators are well-behaved. Thus to gain a deeper understanding of reliability and population classification we will check what happens when one has indicators that reduce reliability. This can be done by adding noise to our measure. We will generate two uncorrelated indicators and substitute x10 and x11 for the indicators x1 and x2. Once the we have introduce some noise to our measure we will estimate a new deprivation score using the two new indicators and dropping x1 and x2. The result is shown below. Then we can apply `omega()` to the new matrix that includes V1 and v1 and excludes x1 and x2. Reliability has drop slightly but enough to raise concerns as both ω and α are below the rules of thumb drawn from a Monte Carlo experiment.

```
#Computing deprivation score with uncorrelated items
Rel_MD_1$ds_ur<-rowSums(Rel_MD_1[,c(3:11)])
Rel_MD_1[1:10,c(16,17)]
```

```
##    hh_members ds
## 1           2  4
## 2           1  0
## 3           1  1
## 4           2  3
## 5           2  1
## 6           1  1
## 7           1  2
## 8           2  1
## 9           2  7
## 10          2  0
```

```
#Now reliability drops
omega_unr_exp<-omega(Rel_MD_1[,c(3:11)])
unrel_uni_exp<-data.frame(omega_exp=omega_unr_exp$omega.tot,
                          alpha=omega_unr_exp$alpha)
unrel_uni_exp
```

```
##   omega_exp    alpha
## 1  0.799405 0.738685
```

What is the impact of introducing the two uncorrelated indicators? From theory is known that losses in reliability affect the consistency of population classification. We can check if this theory holds by looking at the correlation of different rankings that are produce from different measures. For this experiment, first, we will estimate the omega values using different combinations of items (in all case we have the seven items from the reliable measure x1-x9).

```
omega_exp2<-omega(Rel_MD_1[,c(3:9)])
omega_exp3<-omega(Rel_MD_1[,c(1,2,4,5,7,8)])
omega_exp4<-omega(Rel_MD_1[,c(2,3,5,6,8,9)])
```

```
omega_exp5<-omega(Rel_MD_1[,c(1,3,4,6,7,9)])


omegas_exp<-data.frame(omega_exp1=omega_exp1$omega.tot,
                       omega_exp2=omega_exp2$omega.tot,
                       omega_exp3=omega_exp3$omega.tot,
                       omega_exp4=omega_exp4$omega.tot,
                       omega_exp5=omega_exp5$omega.tot,
                       omega_unrel=omega_unr_exp$omega.tot)
```

We then can compare the omega values of each measure. The theory holds for this example. We see that the lowest reliability scale is the one that incorporates V1 and V2. The measures with only seven items have higher reliability. This is a very important lesson as poverty researchers sometimes keep unreliable indicators in their scales and the consequence will be a heavy loss in reliability.

```
t(omegas_exp)
```

```
##                  [,1]
## omega_exp1  0.8599319
## omega_exp2  0.8599319
## omega_exp3  0.8779016
## omega_exp4  0.8543497
## omega_exp5  0.8441886
## omega_unrel 0.7994050
```

The second prediction of reliability theory is that the population orderings are consistent for high reliability values. One way to check this is by estimating the correlation among the different deprivation scores. Again, the theory holds for this simple exercise, the measure with higher ω are highly correlated. The correlation of the unreliable measure seems still high, however, when $\omega < .8$ we could expect to see a classification error $> 5\%$ which might be very worrying when put into perspective. If the poverty rate is 20% and the classification error is 5% it would mean that potentially a 25% of the poor are mistakenly classified (Nájera, 2018).

```
Rel_MD_1$ds_r2<-rowSums(Rel_MD_1[,c(3:9)])
Rel_MD_1$ds_r3<-rowSums(Rel_MD_1[,c(1,2,4,5,7,8)])
Rel_MD_1$ds_r4<-rowSums(Rel_MD_1[,c(2,3,5,6,8,9)])
Rel_MD_1$ds_r5<-rowSums(Rel_MD_1[,c(1,3,4,6,7,9)])

ds.m<-(Rel_MD_1[,c(16:21)])
ds.cor<-cor(ds.m)
ds.cor
```

```
##            hh_members        ds     ds_ur     ds_r2     ds_r3     ds_r4
## hh_members  1.0000000 0.4330385 0.3945498 0.4077791 0.4422686 0.3777748
## ds          0.4330385 1.0000000 0.9272651 0.9684719 0.9764524 0.9523882
## ds_ur       0.3945498 0.9272651 1.0000000 0.9520979 0.8899044 0.8941369
## ds_r2       0.4077791 0.9684719 0.9520979 1.0000000 0.9284786 0.9370200
## ds_r3       0.4422686 0.9764524 0.8899044 0.9284786 1.0000000 0.8875033
## ds_r4       0.3777748 0.9523882 0.8941369 0.9370200 0.8875033 1.0000000
```

3.5.3 Model-based estimation of overall reliability

The ideal workflow in poverty measurement leads to a specification of a model. Different models suggest that poverty is multidimensional and hierarchical (See Section 1.2). Therefore, the interest is in both estimates of reliability: overall and hierarchical omega (ω and ω_h). Both can be estimated from an EFA using the R-package "psych". However, this book is an attempt to encourage poverty researchers to walk toward the

production and assessment of theoretical models. To estimate reliability for a pre-specified model, it is necessary to use Confirmatory Factor Analysis (CFA). Given a pattern loading specification, a CFA will estimate the different parameters of the model. Section 3.3 showed the formulas to estimate both ω and ω_h. Item-factor loadings and the residuals of the model are the key parameters for the estimation of both reliability statistics (See equation (3.6)).

In the following we will show how in both **Mplus** and **R** is possible to estimate ω and ω_h. We will start with R and for this purpose we need the `lavaan` package (Rosseel, 2012). This package comprises a series of functions to estimate different kinds of latent variable models such as measurement and analytic models like Structural Equation Models (SEM). Once the CFA model is fitted with the R-package `lavaan`, the function `omegaFromSem()` of the `psych` R-package can be used to estimate ω and ω_h. However, we will show how this can be done by hand to gain insight of the differences between the two reliability statistics and to operationalise the process using the **Mplus** estimates.

```
#Omega from Sem
library(lavaan)
# We first specify the model
MD_model <- ' h =~ +x1+x2+x3+x4+x5+x6+x7+x8+x9
                F1=~  + x7 + x8 + x9
                F2=~  + x4 + x5 + x6
                F3=~  + x1 + x2 + x3
                h  ~~ 0*F1
                h  ~~ 0*F2
                h  ~~ 0*F3
                F1 ~~ 0*F2
                F2 ~~ 0*F3
                F1 ~~ 0*F3

'
```

To fit the CFA model we will use `sem` function which has been harmonised with the functions `cfa` and `lavaan`. The function requires specifying the measurement model (MD_model), the data, the kind of variables we have (in this case categorical) and we will request standardised loadings with `std.lv=TRUE`.

```
fit <- sem(MD_model, data = Rel_MD_1,
           ordered=c("x1","x2","x3","x4","x5",
                     "x6","x7","x8","x9"),
           std.lv=TRUE)
# The command below is to check the output (We will check this in the
#next section and validity chapter)
#summary(fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

Both ω and ω_h can be manually calculated. There are two main parameters one needs for their computation: factor loadings from the indicators to the overall factor (λ_h), to each dimension (λ_j) and the error. This can be easily extracted from the fit object as follows:

```
lambdas<-as.data.frame(fit@Model@GLIST$lambda)
error<-colSums(fit@Model@GLIST$theta)
```

The then the square of the sum of the loadings (λ_h) and (λ_j) is taken as well as the sum of the error. The we can compute both ω and ω_h using equation (3.6) and (3.7).

```
Slambda_2<-sum(lambdas[1])^2 + sum(lambdas[2])^2 +
           sum(lambdas[3])^2 + sum(lambdas[4])^2
error <- sum(error)

omega_t <- Slambda_2 / (Slambda_2+error)
```

```
omega_h <- sum(lambdas[1])^2 / (Slambda_2+error)
omegamanual<-c(omega_h=omega_h,omega_t=omega_t)
omegamanual
```

```
##   omega_h   omega_t
## 0.8445022 0.9707344
```

Fortunately, there is an R function from the "psych" package that does this for us. Once the model has been fitted, we apply the function `omegaFromSem()` to request the estimates and store the estimates of both ω and ω_h in the `omegasem` object. The results indicate high overall reliability and high reliability after considering the multidimensional features of the scale.

```
omegasem<-omegaFromSem(fit)
omegasem<-c(omega_h=omegasem$omega,
            omega_t=omegasem$omega.tot)
omegasem
```

```
##   omega_h   omega_t
## 0.8446990 0.9707276
```

The R package "mplusAutomation" is an excellent alternative to automate Mplus from R (Hallquist & Wiley, 2018). We can create within R an Mplus object as follows using the function `mplusObject()`. The syntax is the standard Mplus syntax to fit a model. As with `lavaan` we will fit a bi-factor model. We will store the syntax in the object `test`.

```
test <- mplusObject(
TITLE = "Bi-factor model CFA;",
   VARIABLE = "
     NAMES = x1-x9 resources educ_yr occupation class;
     CATEGORICAL = x1-x9;
     USEVARIABLES = x1-x9;",
   ANALYSIS = "ESTIMATOR = wlsmv;
               PROCESS = 4;",

MODEL = "f1 by x1-x3;
  f2 by x4-x6;
  f3 by x7-x9;
  h by x1 x2 x3 x4 x4 x5 x6 x7 x8 x9;
  F1 with F2@0;
  F2 with F3@0;
  F3 with F1@0;
  h with f1@0;
  h with f2@0;
  h with f3@0;",

OUTPUT = "std stdyx;")
```

To write the `test` object as an "*.inp" Mplus syntax file, we will use the function `mplusModeler()`. This function permits estimating the model directly using the option `run`.

```
res <- mplusModeler(test, modelout = "rel_CFA_2.inp",
                    writeData = "never", hashfilename = FALSE,
                    dataout="Rel_MD_data_1_1.dat", run = 1L)
```

```
##
## Running model: rel_CFA_2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "rel_CFA_2.inp"
```

```
## Reading model:  rel_CFA_2.out
```

Once the model has been run, we can import the output using the function `readModels()`. We will explore the full output in the next chapter as for now we will focus in the estimation of ω and ω_h. The factor loadings of the Bi-factor model are stored in a list (parameters). We request the standardised estimates as we did with `lavaan`. We also can request the error from the 'r2 object in the parameters list. Once we have the parameters we need we can proceed as above to estimate the reliability statistics. We see that we could replicate the results from `lavaan`.

```
REL_CFA_2<-readModels(filefilter ="rel_CFA_2")
```

```
## Reading model:  C:/Proyectos Investigacion/PM Book/rel_cfa_2.out
```

```
lambdas<-REL_CFA_2$parameters$std.standardized[1:18,1:3]
error<-REL_CFA_2$parameters$r2[6]

lambda_2<-sum(lambdas[10:18,3])^2 + sum(lambdas[1:3,3])^2 +
          sum(lambdas[4:6,3])^2 + sum(lambdas[7:9,3])^2
error <- sum(error)

omega_t <- lambda_2 / (lambda_2+error)
omega_h <- sum(lambdas[10:18,3])^2 / (lambda_2+error)

omega_t
```

```
## [1] 0.9707333
```

```
omega_h
```

```
## [1] 0.8445348
```

3.5.4 Overall reliability and population orderings

One of the predictions of measurement theory is that reliability leads to consistent population orderings, i.e. poor people will have high deprivation scores and not poor people will have low deprivation scores (see Table). We illustrated this point using the correlation between the different deprivation scores corresponding to diverse levels of overall reliabilities. We can follow up that example by looking at the values of the latent variable for the multidimensional reliable measure (Rel_MD_1). After fitting the CFA model we just can simply use the function `predict()` to obtain the Maximum Likelihood estimates of the latent variable. Then we can merge these values with our data set. The prediction will generate four estimates for the latent variables. The overall factor (h) and the values for the three dimensions.

```
factor_scores<-predict(fit)
Rel_MD_1<-cbind(Rel_MD_1,factor_scores)
head(Rel_MD_1[,c(21:24)])
```

```
##   ds_r4 ds_r5          h          F1
## 1     2     3  0.4475885 -0.81293477
## 2     0     0 -0.6781638 -0.08627376
## 3     0     1 -0.2440620 -0.28822588
## 4     1     2  0.2851351  0.13806885
## 5     0     1 -0.2207057 -0.30252106
## 6     0     1 -0.2207057 -0.30252106
```

To contrast the values of the reliable multidimensional measure with the values of an slitghly less reliable measure we will fit a new model. As in the previous example (Section @ref(#Chapter-3-expoverel)), we will replace the first two indicators x1 and x2 by x10 and x11. Both load into the first factor (f1). The estimates

are stored in a different object (fit_ur) and estimate the factor scores using the `predict()` function. Finally we inspect the values.

```
## We first specify the model
MD_model <- ' h =~ +x10+x11+x3+x4+x5+x6+x7+x8+x9
                F1=~  + x7 + x8 + x9
                F2=~  + x4 + x5 + x6
                F3=~  + x10 + x11 + x3
                h  ~~ 0*F1
                h  ~~ 0*F2
                h  ~~ 0*F3
                F1 ~~ 0*F2
                F2 ~~ 0*F3
                F1 ~~ 0*F3

'

fit_ur <- sem(MD_model, data = Rel_MD_1,
           ordered=c("x10","x11","x3","x4","x5","x6","x7","x8","x9"),
           std.lv=TRUE)
factor_scores_ur<-predict(fit_ur)
colnames(factor_scores_ur)[1:4]<-c("hur","F1ur","F2ur","F3ur")
Rel_MD_1<-cbind(Rel_MD_1,factor_scores_ur)
head(Rel_MD_1[,c(25:28)])
```

```
##           F2         F3         hur       F1ur
## 1 -0.1296229  1.2478407  0.38436469 -0.7563179
## 2 -0.1536510 -0.1720003 -0.59656764 -0.1126966
## 3  0.5290560 -0.3986557 -0.10512455 -0.3763817
## 4 -0.7559100  0.7442695 -0.05228111  0.4723843
## 5 -0.3910445  0.4514504 -0.50856087 -0.1482380
## 6 -0.3910445  0.4514504 -0.59656764 -0.1126966
```

To assess the consistency of both multidimensional scales we will plot the latent factor values by the deprivation score. Figure~3.2 shows that the factor scores are very similar within each deprivation group. For each deprivation score we find very different factor scores, indicating that the deprivation scores is a good measure to rank and split the population according to the severity of deprivation. In contrast, figure~@ref{(fig:fsdesunrel) show that although there is relationship between the deprivation score and factor scores, this relationship is more noisy. Not only there is much more variability within each deprivation group but also there is some overlap. That means that if we use some cut off to split the poor from the not poor based on a deprivation score, we will be more likely to confound both groups. In this case, the mixing of groups is not that dramatic as the scale is still somewhat reliable, but it could be very noisy for less reliable scales.

```
require(ggplot2)
g <- ggplot(Rel_MD_1, aes(as.factor(ds), h))
g + geom_boxplot(varwidth=T) +
    labs(x="Deprivation score. Reliable",
         y="Factor score (Latent variable)") + theme_bw()
```

```
g <- ggplot(Rel_MD_1, aes(as.factor(ds_ur), hur))
g + geom_boxplot(varwidth=T) +
    labs(x="Deprivation score. Unreliable",
         y="Factor score (Latent variable)") + theme_bw()
```

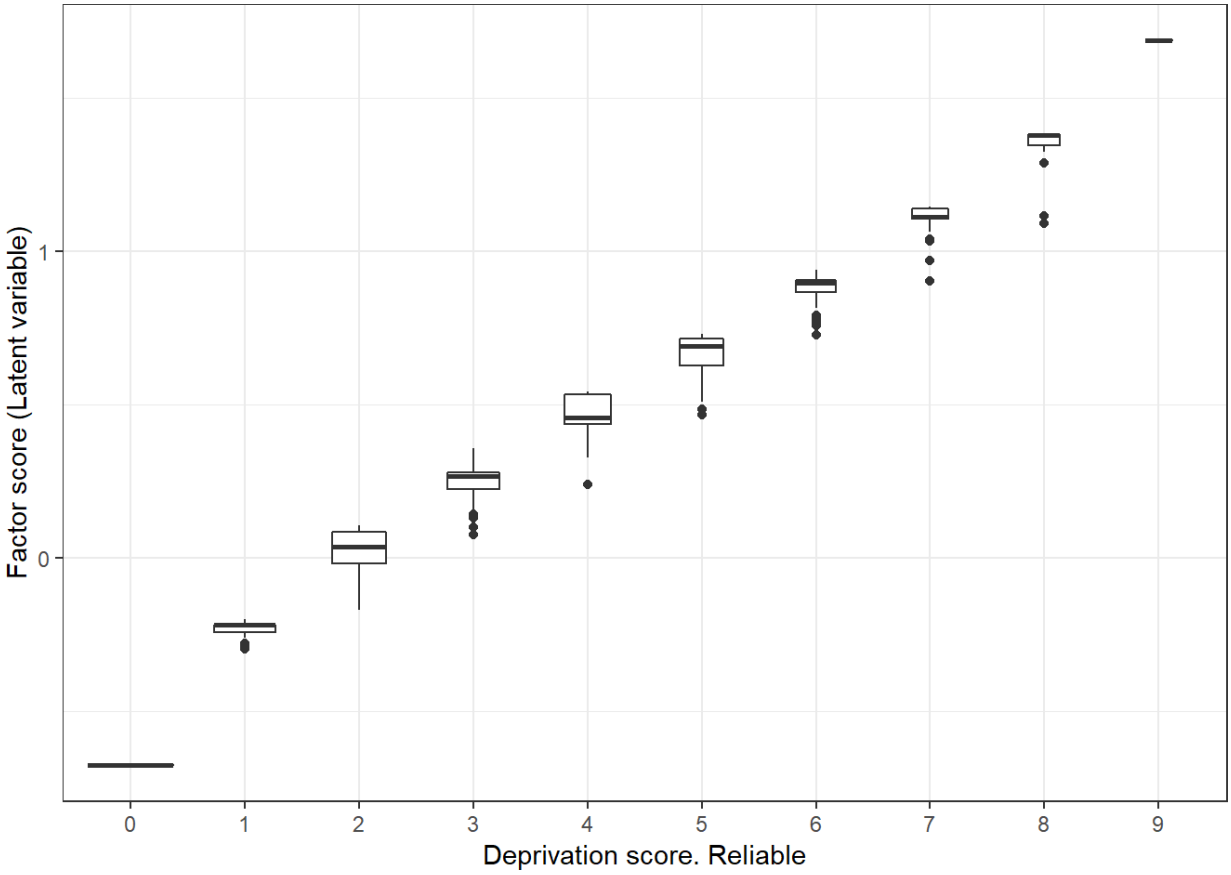

Figure 3.2: Relationship between the deprivation score (x1-x9) and the latent variable score. We appreciate the narrowness of the box plots, indicating good group separation.

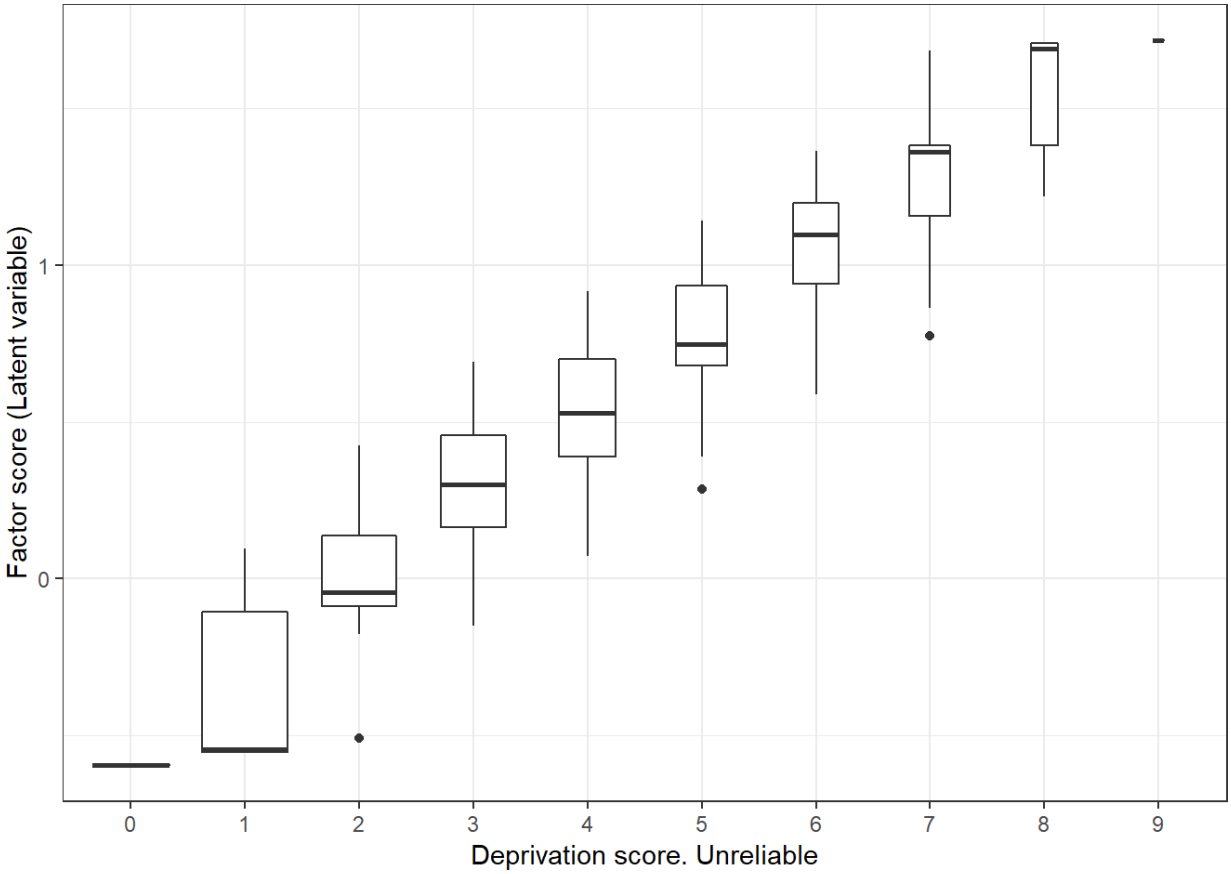

Figure 3.3: Relationship between the deprivation score (x10, x11 and x3-x9) and the latent variable score. There is more variability in this case indicating poor group separation.

3.6 Item-level reliability

Overall reliability is an excellent summary of the quality of an index in that it tells the homogeneity of our scale and its capacity to produce consistent population rankings. In section *ref* we showed that including some uncorrelated items lead to a reduction of reliability and that this has negative implication for consistency across measurements. This section focuses on item reliability, i.e. how specific items contribute positively or negatively to reliability.

Section *ref* reviewed Item Response Theory (IRT), which is a theory of the properties of indicators by putting forward the concepts of discrimination and severity. The standard IRT modelling assummes that scales are unidimensional in that indicators are manifest of a latent trait. However, in parallel to the development of CFA, IRT modelling has incorporated multidimensional models. HOwever, as long as a scale is homogeneous, the bias from a multidimensional IRT and the unidimensional one should not dramatically change our conclusions about the reliability of the items.

To illustrate IRT modelling, we will work with the same data set, which we know that results in a highly homogeneous scale. We will use the data "Rel_MD_1" and the R-package `ltm` which fits different kinds of IRT models. The `ltm()` function fits one, two and three-parameter IRT models. Below we fit a two-parameter IRT model simply by adding the `z1` option so we allow the model to have different slopes, i.e. Two-parameter IRT model. The output shows the difficulty and discrimination coefficients. The difficulty represents the severity of the deprivation on the latent trait. For example, item x1 is less severe than item x3. The second column `Dscrmn` displays the values of the discrimination parameters. All the items have values $> .9$ which is above the suggested threshold by Guio et al. (2016) and Nájera (2018).

```
library(ltm)
rel_irt<-ltm(Rel_MD_1[,c(1:9)] ~ z1)
rel_irt
```

```
##
## Call:
## ltm(formula = Rel_MD_1[, c(1:9)] ~ z1)
##
## Coefficients:
##     Dffclt  Dscrmn
## x1   0.023   2.290
## x2   0.702   2.183
## x3   1.258   2.198
## x4   0.053   2.306
## x5   0.698   2.368
## x6   1.297   2.154
## x7   0.160   2.541
## x8   0.802   2.547
## x9   1.236   2.477
##
## Log.Lik: -20132.92
```

Before checking the estimation with Mplus, we will check what will happen if we include our unreliable items. Items x1 and x2 are replaced by items V1 and V2. As expected the discrimination values are unacceptably low as well as the severity values which are (≥ 3) standard deviations.

```
head(Rel_MD_1[,c(3:11)])
```

```
##   x3 x4 x5 x6 x7 x8 x9 x10 x11
## 1  1  1  0  0  0  0  0   0   0
## 2  0  0  0  0  0  0  0   0   0
## 3  0  1  0  0  0  0  0   0   0
## 4  0  0  0  0  1  0  0   0   0
```

```
## 5  0  0  0  0  0  0  0   1   1
## 6  0  0  0  0  0  0  0   0   0
```

```
rel_irt_2<-ltm(Rel_MD_1[,c(3:11)] ~ z1)
rel_irt_2
```

```
##
## Call:
## ltm(formula = Rel_MD_1[, c(3:11)] ~ z1)
##
## Coefficients:
##      Dffclt  Dscrmn
## x3    1.426   1.651
## x4    0.053   2.371
## x5    0.689   2.488
## x6    1.279   2.245
## x7    0.157   2.817
## x8    0.780   2.840
## x9    1.193   2.804
## x10   5.142   0.133
## x11   4.882   0.135
##
## Log.Lik: -21740.2
```

Similarly we can fit the model using items (x1-x9) in Mplus using the package "mplusAutomation" by creating an object with `mplusObject()`.

```
test <- mplusObject(
TITLE = "IRT model;",
   VARIABLE = "
     NAMES = x1-x9 resources educ_yr occupation hh_size class;
     CATEGORICAL = x1-x9;
     USEVARIABLES = x1-x9;",
   ANALYSIS = "ESTIMATOR = ml;
               PROCESS = 4;",

MODEL = "h by x1* x2-x9;
         h@1;")
```

Then the model is fitted using `mplusModeler()`. We will name our Mplus script as "rel_IRT_1.inp" and we will be using the same data as before ("Rel_MD_data_1_1.dat") and we request Mplus to run the model directly from the script with (`run`).

```
res <- mplusModeler(test, modelout = "rel_IRT_1.inp",
                    writeData = "never", hashfilename = FALSE,
                    dataout="Rel_MD_data_1_1.dat", run = 1L)
```

Now we read the result of our model with `readModels()`.

```
REL_IRT_1<-readModels(filefilter ="rel_IRT_1")
```

The `readModels()` does an excellent job in extracting and ordering the Mplus output. It puts all the relevant result in lists. We can see that the Mplus estimates are very similar to those obtained from the `ltm` package.

```
rel_irt<-REL_IRT_1$parameters$irt.parameterization
rel_irt<-rel_irt[1:18,]
rel_irt<-data.frame(a=rel_irt$est[1:9],b=rel_irt$est[10:18])
rel_irt
```

```
##       a     b
## 1 2.290 0.024
## 2 2.192 0.700
## 3 2.207 1.254
## 4 2.312 0.054
## 5 2.383 0.696
## 6 2.172 1.291
## 7 2.548 0.161
## 8 2.561 0.800
## 9 2.498 1.231
```

3.7 Multidimensional item-reliability evaluation

A multidimensional IRT model is just a CFA model with categorical indicators. One way to assess the item-level reliability is by looking at the loadings from a CFA model. We can fit a higher-order factor model (equivalent to the bi-factor model above) to assess the value of the loadings and and look at the R^2 values of each indicator (which is just λhj^2). $R^2 \leq .25$ are equivalent to $\lambda hj^2 \leq .5$, which are often used as cut offs of unacceptably low loadings. We see that all these items are highly reliable not only with regard the higher order factor but also in terms of each dimension, measured by each factor loading.

```
 MD_model <- ' f1  =~ x1 + x2 + x3
              f2 =~ x4 + x5 + x6
              f3   =~ x7 + x8 + x9
                h =~ f1 + f2 + f3
 '

fit <- sem(MD_model, data = Rel_MD_1,
           ordered=c("x1","x2","x3","x4","x5",
                     "x6","x7","x8","x9"))
inspect(fit,what="std")$lambda
```

```
##       f1    f2    f3 h
## x1 0.924 0.000 0.000 0
## x2 0.888 0.000 0.000 0
## x3 0.873 0.000 0.000 0
## x4 0.000 0.929 0.000 0
## x5 0.000 0.917 0.000 0
## x6 0.000 0.866 0.000 0
## x7 0.000 0.000 0.947 0
## x8 0.000 0.000 0.916 0
## x9 0.000 0.000 0.894 0
```

3.7.1 Item-reliability and monotonicity

Low loadings are an indication that the indicator is not a manifest variable of the underlying construct. That is, that changes in poverty do not mirror changes in deprivation. Section~**??** suggested that there is a relationship between item-reliability and the monotonicity axiom. Nájera (n.d.) shows that indeed low loadings approximately $\leq .5$ lead to violations of the strong monotonicity axiom, i.e. a reduction in poverty does not reflect an improvement in the achievement matrix. Weak monotonicity is violated when an improvement in poverty results in an increase of deprivation, this would happen when the factor loadings are negative, for example.

We will fit a higher order model using again the unreliable items (x10 and x11) instead of x1 and x2. Of course, it is possible to use either the loadings of the R^2 values (these can be obtained with the `summary()` function. We see again, as in the IRT analysis that both items have unacceptably low values. These items

should be dropped from the scale as it inclusion introduces noise to our measure. That would imply dropping the first dimension in the absence of alternative indicators.

```
 MD_model <- ' f1  =~ x10 + x11 + x3
               f2 =~ x4 + x5 + x6
               f3   =~ x7 + x8 + x9
                 h =~ f1 + f2 + f3
 '

fit <- sem(MD_model, data = Rel_MD_1,
           ordered=c("x10","x11","x3","x4","x5",
                     "x6","x7","x8","x9"))
```

```
## Warning in lav_object_post_check(object): lavaan WARNING: some estimated ov
## variances are negative
```

```
inspect(fit,what="std")$lambda
```

```
##        f1    f2    f3 h
## x10 0.114 0.000 0.000 0
## x11 0.116 0.000 0.000 0
## x3  1.036 0.000 0.000 0
## x4  0.000 0.924 0.000 0
## x5  0.000 0.920 0.000 0
## x6  0.000 0.868 0.000 0
## x7  0.000 0.000 0.947 0
## x8  0.000 0.000 0.914 0
## x9  0.000 0.000 0.898 0
```

3.8 Real data example

We will use the Mexican data set "Mex_pobreza_14.dat". This data set contains a subset of the deprivation indicators utlised to measure multidimensional poverty in Mexico. The official measure has two domains: income and social rights. The social rights domain has five dimensions: essential services, housing, food deprivation, social security and education. Some of this dimensions are measured with few indicators, like education and social security. This poses limitations to fit an identified model. Hence, we will use a reduced version of the model comprising three dimensions: essential services, housing and food deprivation.

The Mexican poverty data comes from a nationally representative complex survey. We will use the package "survey" (Lumley, 2016). This is a comprehensive R-package to analyse survey data. We strongly advice readers to check Lumley (2011) book on complex surveys to get a depth insight on complex sampling and the use of Lumley's excellent package. To produce design deprivation rates estimates for each of the 14 items we need to specify few things. First we need to identify the sampling weights and the primary sampling units (PSU). The (`options()`) function is to prevent errors as sometimes there is one household per PSU. Once we have set up the sampling design we can estimate the deprivation rates with the `svymean()` function -we round the percentages for simplicity-. We can see than deprivation in the housing dimension items is rather low. Essential services present higher deprivation rates but electricity is very low. We see that the food deprivation items have the higher rates on average.

```
library(haven)
Mex_D<-read_dta("pobreza_14.dta")

cols <- c("icv_muros", "icv_techos", "icv_pisos", "icv_hac",
          "isb_agua","isb_dren", "isb_luz", "isb_combus",
          "ic_sbv", "ia_1ad", "ia_2ad", "ia_3ad", "ia_4ad",
          "ia_5ad",  "ia_6ad")
```

```
library(survey)
options(scipen=999, survey.lonely.psu="adjust")
des <- svydesign(data=Mex_D, id=~1, CLUSTER=~psu, weights=~weight)
propr <- data.frame(svymean(Mex_D[, cols],des,na.rm=T))
propr <- round(propr*100,1)
propr
```

```
##            mean  SE
## icv_muros   1.7 0.1
## icv_techos  1.6 0.1
## icv_pisos   3.0 0.1
## icv_hac     5.6 0.1
## isb_agua    7.7 0.1
## isb_dren    7.5 0.1
## isb_luz     0.8 0.0
## isb_combus 12.0 0.2
## ic_sbv     19.6 0.2
## ia_1ad     33.3 0.3
## ia_2ad     15.9 0.2
## ia_3ad     24.9 0.2
## ia_4ad     14.0 0.2
## ia_5ad     16.4 0.2
## ia_6ad     12.3 0.2
```

Now the we are familiarised with the 14 deprivation indicators we can proceed to estimate the reliability statistics. We will fit the model in Mplus as we will incorporate the survey design in the estimation of the parameters -it is possible to do so with the `lavaan.survey()` too-. We first create our Mplus script (rel_CFA_mex.inp) to fit the bi-factor model. Following the theoretical model for this data, the model has three dimensions (housing (f1), essential services (f2) and food deprivaton (f3)) and one higher-order factor (h). Once the model has been fitted we will store the output in the object called `REL\_CFA\_mex`.

```
test <- mplusObject(
TITLE = "Bi-factor model CFA;",
   VARIABLE = "
     NAMES = proyecto folioviv foliohog icv_muros icv_techos
             icv_pisos icv_hac isb_agua isb_dren isb_luz isb_combus
             ic_sbv ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad
             ia_7men ia_8men ia_9men ia_10men ia_11men ia_12men
             tv_dep radio_dep fridge_dep
             washingmach_dep compu_dep inter_dep psu weight
             rururb tot_integ durables educ_hh;
MISSING=.;
     CATEGORICAL = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
     USEVARIABLES = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;

WEIGHT=weight;
cluster = psu;",

   ANALYSIS = "TYPE = complex;
```

```
ESTIMATOR = wlsmv;
PROCESS = 4;",

MODEL = "f1 by icv_muros icv_techos icv_pisos icv_hac;
  f2 by isb_agua
        isb_dren isb_luz isb_combus;
  f3 by ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
  h  by icv_muros icv_techos icv_pisos icv_hac isb_agua
        isb_dren isb_luz isb_combus ia_1ad ia_2ad
        ia_3ad ia_4ad ia_5ad ia_6ad;
  F1 with F2@0;
  F2 with F3@0;
  F3 with F1@0;
  h with f1@0;
  h with f2@0;
  h with f3@0;",

OUTPUT = "std stdyx;")

mplusModeler(test, modelout = "rel_CFA_mex.inp",
                    writeData = "never", hashfilename = FALSE,
                    dataout="Mex_pobreza_14.dat", run = 1L)
REL_CFA_mex<-readModels("rel_CFA_mex.out")
```

We then can estimate the overall reliability statistics; both ω and ${\omega_h}^2$. The reliability of our measure is very high under both statistics and both are above the recommended thresholds (Nájera, 2018). To estimate the reliability measures we will use the same approach we followed in the previous section. First, we will obtain the factor loadings (f's and h) and the errors for each indicator. Then we we will take the square of the sum of the lambdas (f1, f2, f3 and h) and the error sum. ω and ω_h then can be calculated using formulas from equations (3.6) and @ref(eq:omegah}).

The values of both ω and ω_h are high .97 and .81, respectively. These figures suggest that this multidimensional scale is homogeneous but has multidimensional features ($\omega_h < \omega$).

```
lambdas<-REL_CFA_mex$parameters$std.standardized[1:28,1:3]
error<-REL_CFA_mex$parameters$r2[6]

lambda_2<-sum(lambdas[10:28,3])^2 + sum(lambdas[1:4,3])^2 +
          sum(lambdas[5:8,3])^2 + sum(lambdas[9:14,3])^2
error <- sum(error)

omega_t <- lambda_2 / (lambda_2+error)
omega_h <- sum(lambdas[10:28,3])^2 / (lambda_2+error)

omega_t
```

```
## [1] 0.9730641
```

```
omega_h
```

```
## [1] 0.8058398
```

[2]Prior the estimation of the reliability statistics is vital to inspect the fit of the model as a poor model will not be useful to estimate omega. We will discuss in the next chapter (Validity) the meaning of these statistics. At this point we will focus on the fact that a poor model fit invariably leads to poor estimates of reliability. There is no point in estimating reliability of a scale that makes no sense at all. To assess the fit of the model we look at four statistics: χ^2, TLI, CFI and RMSEA. For this model the relative statistics of fit look fine and we have some certainties about the model we fit.

3.8.1 Item-level reliability

Once we have assessed the overall reliability of the Mexican index, we can assess item-reliability by fitting a higher-order factor model and checking the value of the factor loadings. We just simply need to rewrite our model (rel_CFA_mex2.inp) to represent a structure were the dimensions load into the higher order factor. This is simply done by specifying that h is measured by the three dimensions (f1, f2 and f3).

```
test <- mplusObject(
TITLE = "CFA higher order model CFA;",
   VARIABLE = "
     NAMES = proyecto folioviv foliohog icv_muros icv_techos
             icv_pisos icv_hac isb_agua isb_dren isb_luz isb_combus
             ic_sbv ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad
             ia_7men ia_8men ia_9men ia_10men ia_11men ia_12men
             tv_dep radio_dep fridge_dep
             washingmach_dep compu_dep inter_dep psu weight
            rururb tot_integ;
MISSING=.;
     CATEGORICAL = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
     USEVARIABLES = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;

WEIGHT=weight;
cluster = psu;",

   ANALYSIS = "TYPE = complex;

ESTIMATOR = wlsmv;
PROCESS = 4;",

MODEL = "f1 by icv_muros icv_techos icv_pisos icv_hac;
  f2 by isb_agua
        isb_dren isb_luz isb_combus;
  f3 by ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
  h by f1 f2 f3;",

OUTPUT = "std stdyx;")

mplusModeler(test, modelout = "rel_CFA_mex2.inp",
                    writeData = "never", hashfilename = FALSE,
                    dataout="Mex_pobreza_14.dat", run = 1L)
```

Once the model has been fitted we can request the standardised factor loadings and inspect its values using the `readModels()`. To facilitate our interpretation we can plot the standardised loadings (Figure 3.4. We see that all the indicators have very high loadings and above the suggested threshold ($\lambda_i j > .5$). That means that these indicators discriminate well and are reliable manifests of each dimension. We see, nonetheless, that the indicators of the housing dimension tend to have low values. This could be an indication that these indicators are losing discriminatory power due to changes in living standards.

```
REL_CFA_mex2<-readModels("rel_CFA_mex2.out")
modelParams<- REL_CFA_mex2$parameters$std.standardized[1:14,]
modelParams <- subset(modelParams, select=c("paramHeader", "param", "est", "se"))
```

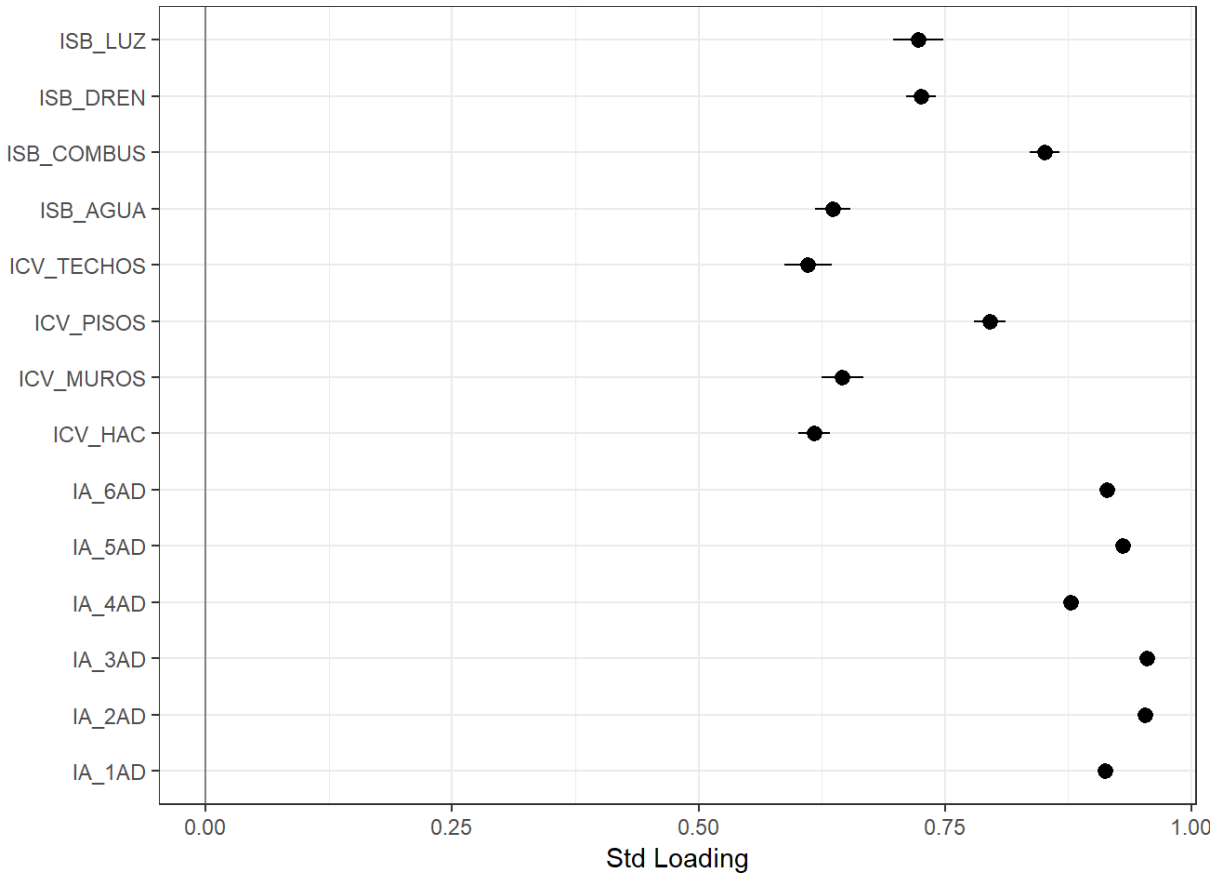

Figure 3.4: This plot shows the standardised values of the loadings for the 14 items.

```
library(ggplot2)
limits <- aes(ymax = est + se, ymin=est - se)
 ggplot(modelParams, aes(x=param, y=est)) + geom_pointrange(limits) +scale_x_discrete("") +
   geom_hline(yintercept=0, color="grey50") + theme_bw() +ylab("Std Loading") + coord_flip()
```

Chapter 4

Validity in poverty measurement

Abstract

This chapter focuses on the theory and implementation of validity. An intuitive explanation is provided about the relationship between reliability and validity and then the different definitions of validity are reviewed. The chapter then uses simulated data to illustrate how construct and criterion validity can be investigated using R and Mplus. The chapter finalised by looking at a real-data example.

4.1 Intuition to the concept of validity

We have seen so far that reliability is homogeneity in measurement and it means the capacity of a measure to reproduce the ranking of a population under changing conditions. Reliability, thus, 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.

Imagine that we know the standards of living of two subject 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. is more likely to be poor than the second. 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.

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 drink 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 behaviours. Then we could compare our measurement to hers to assess the precision of our instrument.

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. The way forward has been an unified framework of validity which looks at the extent to which the existence evidence on a scale supports the intended interpretation of test scores for the proposed use (AERA, APA and NCME, 2014).

4.2 Theory of validity

Classical test theory (CTT) proposes that reliability is the maximum possible validity of a scale. Reliability is affected by both systematic and random error but systematic error only affect validity. How this is possible

according to CTT? The observed score is just the combination of the true score plus error. Validity is a function of systematic error (i.e. constant deviations from the construct of interest) 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. In CTT validity is formulated as follows:

$$V = \frac{\sigma^2_{CI}}{\sigma^2_{observed}} \tag{4.1}$$

The problem with this formulation is that it was little practical usage as the problem in question is knowing σ^2_{CI}. The best approximation for this notion of validity consisted in focusin on the predictive capacity of a scale. Nonetheless, this approach to validity changed after the 1950s. Bandalos (2018) provides and overview of how the *standards* have both discussed and expanded the definition of validity 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 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. For example, in Townsend's theory there are five main types of resources. Drawing upon, Townsend (1979), criterion validity has been used in poverty measurement by Guio et al. (2012) for the production of the European deprivation index and by Gordon (2010) in his proposal for the Mexican multidimensional measure. Similarly, Nandy & Pomati (2015) used criterion validity to assess their proposed index for Benin.

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, for example, in poverty research it could be acute poverty or housing and facilities deprivation. In other settings, 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. Because the Mexican law represents the will of the people, this gives a content validity to the measure (Gordon, 2010). This, nonetheless, does not means that the law will lead to a valid scale. It means that the one should assess validity in accordance to the definition, i.e. examine whether the dimensions and indicators lead to a reliable and valid measure.

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, 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. The use of different kinds of good data will enhance the theory that gives content to a concept. **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. In other words, how sensible a the contents of a poverty index seem to the poor and the not poor. 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 et al. (2006); Gordon (2018)). This project follows the ideal work flow production of a poverty measure (see 2.2) in that the concept of poverty has a theory that defines it, the questionnaire is first calibrated with qualitative work (face validity), and then a survey questionnaire is developed with the explicit purpose of measuring multidimensional poverty.

However, as discussed in Figure 2.1, in practice poverty researchers work with the data they already have and content validity is constrain and face validity is ignored. Furthermore, in many cases, criterion validity might not be available due to the fact that there is no a clear priority about the aspects of a concept that should be targeted in a measurement exercise. Furthermore, it might be the case that there is no clear predictor to

conduct criterion validity. For example, when there is no agreement about the causes and correlate variables of poverty. Cronbach & 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, at first, was seeing as the last resource but in the contemporary literature is no longer the case. 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.

4.3 Methods for the analysis validity

4.3.1 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 book is not about the different explanations of poverty but requires an illustration of the importance of having a theory for validity. 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 et al., 2006). Poverty theories are crudely classified into structural and individual-centred theories. Peter Townsend (1993) and Townsend (1979) provide and overview of these frameworks. To illustrate how a theory of poverty relates to validity we will draw upon Townsend's theory as it is one of the few frameworks that links an explanation of poverty with a cogent theory for its measurement.

Townsend predicts that the lack of command of different kinds of resources lead to deprivation. This means that poverty should be related with different expressions of command of resources such as position in the labour market, education attainment, social class, etc. Poverty leads to exclusion and is said to be a good predictor of ill-health, therefore, poverty should be correlated with a measure of health status too. This means that the type of the employment should be correlated with our index of poverty, for example.

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. He used income as a measure of resources given that the Mexican Income and Expenditure Survey lacked a validator (we reflected on the problem of data production for poverty measurement see section). He did fit a Generalized Linear Model (GLM) using a binary variable (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. In Gordon (2010)'s example the validator is far from ideal but illustrates the idea of criterion validation. In figure 4.1 it is proposed a slightly better validator such as the position in the labour market. International occupation scales that aim to measure socio-economic position could be useful for this purpose (Ganzeboom & Treiman, 1996). There are other alternatives like using subjective indicators of well-being or self-assessments of health status as those used by Guio et al. (2012).

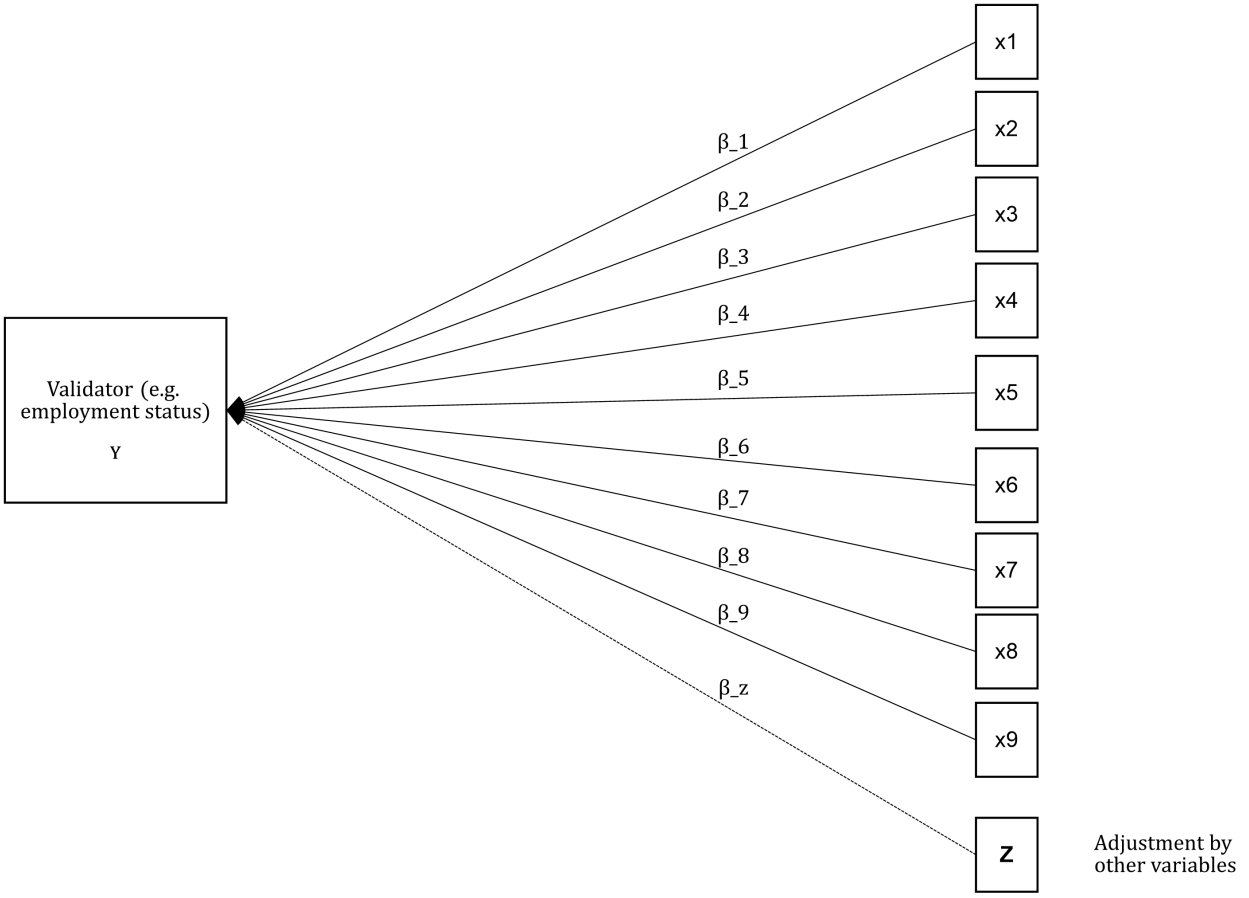

Figure 4.1: This is a visual representation of @Gordon2010 criterion validation. Here income poverty is replaced by a measure of socio-economic position like employment status.

The advantage of using a unified framework such as measurement theory is that it is possible to further specifying a criterion validation model in terms of a Confirmatory Factor Model plus and explanatory model (Structural Equation Modelling, SEM). In the latent variable literature, these kind of models are known as MIMIC models or Multiple Indicator, Multiple Cause. Figure 4.2 shows a visual representation of the criterion validation of the unidimensional model previously shown. In this case, there is a new path from the validator (Y) toward the latent variable. The model is adjusted by a series of covariates (Z). In this model the expectation would be to see that β_y to be associated with the factor (poverty) in a sensible way. If the metric of the factor (often standardised with mean zero and variance equal to one), tells that higher values denote higher severity, then we should expect that people with non-skilled jobs to be associated with positive factor values.

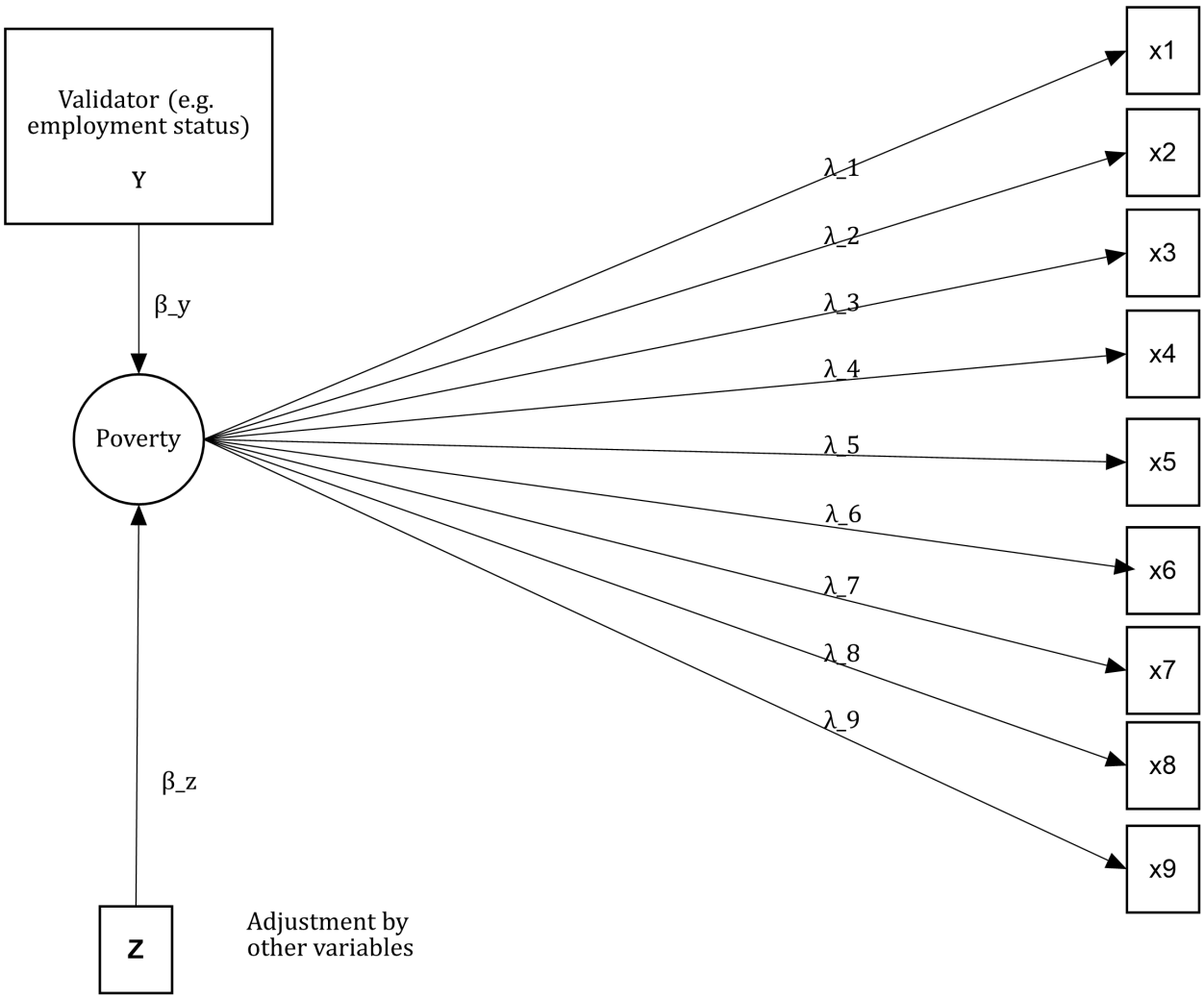

Figure 4.2: This is a visual representation of a MIMIC criterion validation of a unidimensional or null model

We could easily extend this example of criterion validation for a simplified version of the Townsend model. Figure 4.3 shows that poverty is predicted by a validator and some auxiliary variables (Z). The expectation is to find a predictive relationship between Y and the latent variable. The rest is just Townsend's measurement model of poverty. In this reduced version, only the two main dimensions are presented. In the next section is discussed that having just two dimensions leads to some identification issues for the empirical analysis.

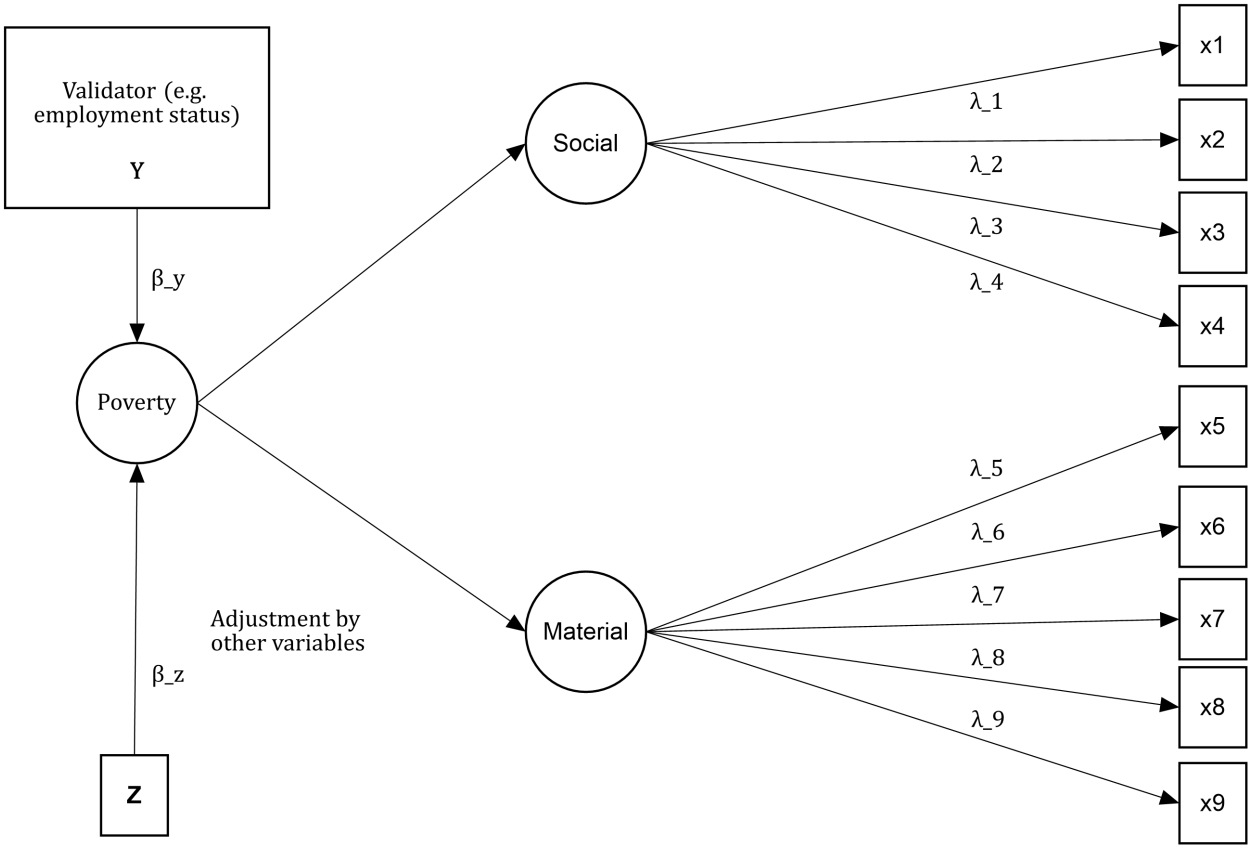

Figure 4.3: This is a visual representation of a MIMIC criterion validation of a reduced version of the theoretical model of Townsend

4.3.2 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. To ilustrate 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. These models have evolved to such extent that the most powerful factor model could be used to test in one model a number of hypothesis. There are two main ways to conduct factor analysis: exploratory and confirmatory. This book puts emphasis on the second kind as it soughs to encourage the development of scales based on theory and not on what the available data says. The label "Confirmatory" is ambitious in that it suggest that we confirm that our model is right. This, of course, is

never possible. The best we can do is to assess whether the model is not a bad one, which does not necessarily means that is the correct one.

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. This is no different from any experiment where given some assumptions, researchers compute if their model of reality is matched by observation. CFA models aim to assess if the presumed model of poverty seems to hold given the data, i.e. whether there is any indication that there are three dimensions, the indicators seem to relate to these dimensions and the contribution of the indicators is equally important or not within dimensions.

To explain the theory of CFA models is necessary to bring back equations (2.2) and (2.3).

$$x_{ij} = \lambda_{ij}\eta_j + \varepsilon_i j \tag{4.2}$$

$$\eta_j = \gamma_j \zeta + \xi \tag{4.3}$$

These equations represent a hierarchical Confirmatory Factor model. These make our measurement model testable using a method that was developed for such purpose. This model will tell us: if the three dimensions η_j $(j = 1, 2, 3)$ is an adequate representation of poverty. It will also tell us if the indicators are manifest (λ_{ij}) of the presumed dimensions, and whether the loadings are equal or not within dimensions.

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. Maximum Likelihood (ML) is one of the preferred methods to estimate the minimizing function F_{ML} (see p. 72 and 73 for an explanation)(Brown, 2006). There are, nonetheless, several estimating procedures that are more or less adequate depending on the nature of the data. One of the most useful and adequate for the kind of data in poverty measurement (categorical data with large samples) is robust weighted least squares (WLSMV) as it is faster than ML and is asymptotic distribution free. F_{ML} is very useful because it provides standard errors (SEs) of the estimates but also because it can be used for the calculation of several indices of goodness-of-fit which tell how poor or good the model is.

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 tell 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 $S = \Sigma$.

A relative index of goodness-of-fit is root mean square error of approximation (RMSEA) (Steiger, 1980). This index looks at the extent to which a model is a reasonable approximation in the population. This index is sensible to the number of parameters in the model but insensitive to sample size.

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.

Several Monte Carlo studies have been conducted to assess the behaviour of these indices (Bentler, 2007; Browne, Cudeck, & others, 1993; Hu & Bentler, 1999; Rigdon, 1996). From these studies it has been possible to have an approximation to the values of the indices that often indicate a good fit. These values are summarised as follows:

Table 4.1: Summary of the suggested cut off for the goodness-of-fit statistics. The values of RMSEA, CFI and TLI need to be taken as an approximation.

Index	Range values	Poor model fit rule
χ^2	p-values 1-0	$p < .05$
$RMSEA$	p-values 1-0	$p > .06$
CFI	$1 - 0$	$< .95$
TLI	$1 - 0$	$< .95$

Factor loadings are often thought as a measure of item-reliability (see Section @ref()). So how does the factor loading values fit in a validity analysis? There is no consensus about threating factor loadings as measures of item validity. Only and only if the measure is proven to be valid in some way, it is possible to frame item loadings in terms of validity. In such a context, 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. The next section shows how these tests work but the idea is to assess the extent to which $\lambda_{11} = \lambda_{21} = \lambda_{31}$, for example for three items in dimension $j = 1$.

4.4 Validity assessment

4.4.1 Criterion Validity

Criterion or predictive validity holds when there is a correlation between an scale and an alternative measure on the cause or effects of the construct of interest. In poverty research, this kind of validation has been used in the empirical literature (Gordon, 2010; Guio et al., 2012; Nandy & Pomati, 2015). We will again use our simulated multidimensional measure to illustrate how a validation exercise can be undertaken and to underline some issue researchers might find in practice.

Fitting a regression model to assess the relationship between a proposed index and an alternative measure is a common approach to assess predictive validity. To illustrate how this kind of validation works, we will use the simulated data ("Rel_MD_data_1_1.dat"). This data set contains the nine manifest variables (x1-x9) plus the two unreliable indicators(x10-x11). Three variables were simulated as alternative measures. One is a "perfect" measure of the resources available for each household in the sample. So in principle, this measure ranks the households according to their potential to fulfil their needs. The measure is expressed in monetary terms to facilitate the interpretation. Education years of the household head and occupation (skill scale) are two predictors of the living standards of the households. These two variables reflect the often common case where the survey was not designed with a validator in mind. We will use the variable "hh_members" to adjust the estimates.

```
library(plyr)
Rel_MD_1<-read.table("Rel_MD_data_1_1.dat")
Rel_MD_1$ds<-rowSums(Rel_MD_1[,c(1:9)])
colnames(Rel_MD_1)<-c("x1","x2","x3","x4","x5","x6",
                      "x7","x8","x9","x10","x11",
                      "resources","educ_yr","occupation",
                      "hh_members","class","ds")
Rel_MD_1[1:5,1:11]
```

```
##   x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11
## 1  1  1  1  1  0  0  0  0  0   0   0
## 2  0  0  0  0  0  0  0  0  0   0   0
## 3  0  0  0  1  0  0  0  0  0   0   0
## 4  1  1  0  0  0  0  1  0  0   0   0
```

```
## 5  1  0  0  0  0  0  0  0  0   1   1
```

```
Rel_MD_1[1:5,12:15]
```

```
##   resources educ_yr occupation hh_members
## 1  3276.687       6          4          5
## 2  7508.982      15          2          1
## 3  7183.707       8          5          2
## 4  1574.356       6          2          7
## 5  2210.297       9          5          4
```

One way to conduct the validation analysis consists in estimating the association between the manifest variables of our index with the validator. This can be simply done by fitting a series of regression models. Because deprivations are binary variables, we need to use a Generalised Linear Model (GLM) with the appropriate distribution. Relative Risk Ratios (RRR) are easier to interpret, so we will fit a Poisson model with log link to obtain the RRRs. Of course, there is no problem in estimating odd-ratios as here we are interested in looking at the association between variables.

In total we have 11 dependent variables (x1-x11) and, thus 11 models. In principle, x1-x11 resulted unreliable and should have been dropped from the analysis but we will keep them just to discuss some connections between reliability and validity. We will create a simple function `lms()` below to loop across the deprivation indicators. We will also transform the resources to get a more sensible metric.

```
Rel_MD_1$resources<-Rel_MD_1$resources*.01

lms<-function(index)
{
  fit<-glm(Rel_MD_1[,index] ~ Rel_MD_1$resources +
                              Rel_MD_1$hh_members,
           family=poisson(link="log"))
  exp(cbind(OR = coef(fit), confint(fit)))
}

coefs<-lapply(1:11,lms)

coefs[[1]]
```

We could check each of the outputs in list `coefs` but it is easier to plot the RRRs of resources for each one of the 11 variables. We will not show the code here but one could just simply extract the coefficients and use `ggplot2()` to produce the graph. The coefficients are displayed with 95% confidence intervals in plot~4.4[1]. The null hypothesis in this model is that there is no relationship between resources and deprivation. For items x1-x9 we see that the difference seems to be different from zero and that the estimates are likely to be less than one. This suggests the higher the resources and lower the chances of being deprived. This is in line with our expectation. For items x10 and x11, however, we found no relationship at all. This is an indication that both items are unreliable and invalid. This reinforces our previous suspicion that these two items are not useful to measure poverty.

```
coefs<-lapply(coefs, function(x) unlist(x[2,]))
coefs<- as.matrix(matrix(unlist(coefs), nrow=length(coefs), byrow=T))
coefs<-data.frame(rbind(coefs[,c(1,2,3)]))

coefs$item <- rep(c("x1","x2","x3","x4","x5","x6",
                    "x7","x8","x9","x10","x11"),1)
coefs$var<-c(rep("Resources (*100)", 11))
```

[1]Here we are using classic or frequentist statistics. There are many problems around the use of p-values. We will be careful in the interpretation as the kind of test we run here is very conservative, i.e. the association is zero. We are not assessing whether is positive or negative. But we should do it in future editions.

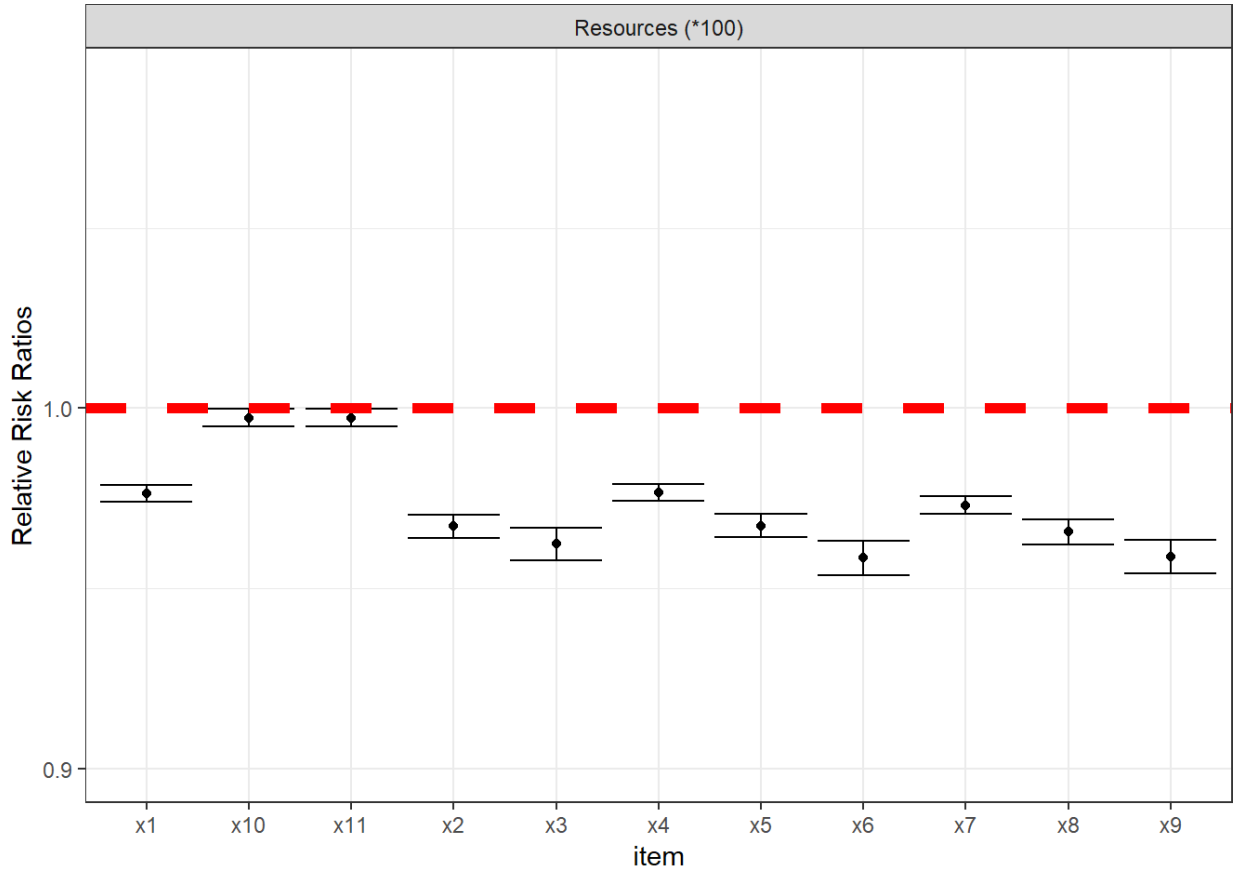

Figure 4.4: This plot shows the Relative Risk Ratios for the resources variable, adjusted by the household size. Having more resources reduces the risk of being deprived of the item x, as expected.

```
coefs
```

We can simply plot the coefficients of each variable using the object `coefs` and `ggplot2()` as follows:

```
p<- ggplot(coefs, aes(x=item,y=X1)) + geom_point() +
  geom_errorbar(aes(ymin=X2, ymax=X3)) +
 theme_bw() + scale_y_continuous(trans = 'log10', limits = c(.9, 1.1))
p + facet_grid(. ~ var) + labs(y="Relative Risk Ratios") + geom_hline(yintercept=1, linetype="dashed",
                color = "red", size=2)
```

Now we will go through the case of the lack of a validator. Most of the time researchers will lack a validator that was designed a priori. In these circumstances researchers need to use variables that predict poverty. Education attainment of the household head and occupation status are one of the two best predictors of poverty. We will rewrite our `lms()` function to fit a series of models using both education and occupation. All models adjusted by the household size. Again we will fit a GLM to obtain relative risks.

```
lms<-function(index)
{
  fit<-glm(Rel_MD_1[,index] ~ Rel_MD_1$occupation +
                              Rel_MD_1$educ_yr +
                              Rel_MD_1$hh_members,
           family=poisson(link="log"))
  exp(cbind(OR = coef(fit), confint(fit)))
```

```
}

coefs<-lapply(1:11,lms)
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

```
coefs[[1]]
```

```
##                             OR     2.5 %    97.5 %
## (Intercept)         0.6866791 0.5503763 0.8554189
## Rel_MD_1$occupation 1.0531859 1.0377362 1.0688189
## Rel_MD_1$educ_yr    0.9206710 0.9051449 0.9364577
## Rel_MD_1$hh_members 1.0163837 0.9939741 1.0391996
```

```
coefs<-lapply(coefs, function(x) unlist(x[2:3,]))
coefs<- as.matrix(matrix(unlist(coefs), nrow=length(coefs), byrow=T))
coefs<-data.frame(rbind(coefs[,c(1,3,5)],coefs[,c(2,4,6)]))

coefs$item <- rep(c("x1","x2","x3","x4","x5","x6",
                    "x7","x8","x9","x10","x11"),2)
coefs$var<-c(rep("Occupation (Skill) scale", 11), rep("Education years", 11))
coefs
```

```
##           X1        X2        X3 item                       var
## 1  1.0531859 1.0377362 1.0688189   x1 Occupation (Skill) scale
## 2  1.0691741 1.0489816 1.0896804   x2 Occupation (Skill) scale
## 3  1.0781765 1.0510322 1.1058950   x3 Occupation (Skill) scale
## 4  1.0590273 1.0433546 1.0748872   x4 Occupation (Skill) scale
## 5  1.0721107 1.0517452 1.0927967   x5 Occupation (Skill) scale
## 6  1.0993761 1.0713526 1.1280058   x6 Occupation (Skill) scale
## 7  1.0610757 1.0448192 1.0775352   x7 Occupation (Skill) scale
## 8  1.0728834 1.0512460 1.0948827   x8 Occupation (Skill) scale
## 9  1.1084601 1.0802115 1.1373136   x9 Occupation (Skill) scale
## 10 1.0001156 0.9817638 1.0187286  x10 Occupation (Skill) scale
## 11 1.0087886 0.9905182 1.0273171  x11 Occupation (Skill) scale
## 12 0.9206710 0.9051449 0.9364577   x1          Education years
## 13 0.8977545 0.8780443 0.9178932   x2          Education years
## 14 0.8758993 0.8500184 0.9025356   x3          Education years
## 15 0.9311522 0.9153057 0.9472680   x4          Education years
## 16 0.9052537 0.8851807 0.9257679   x5          Education years
## 17 0.8979686 0.8708921 0.9258545   x6          Education years
## 18 0.9150086 0.8987985 0.9315038   x7          Education years
## 19 0.9029371 0.8816859 0.9246844   x8          Education years
## 20 0.8974289 0.8704803 0.9251801   x9          Education years
## 21 0.9848866 0.9650072 1.0051781  x10          Education years
```

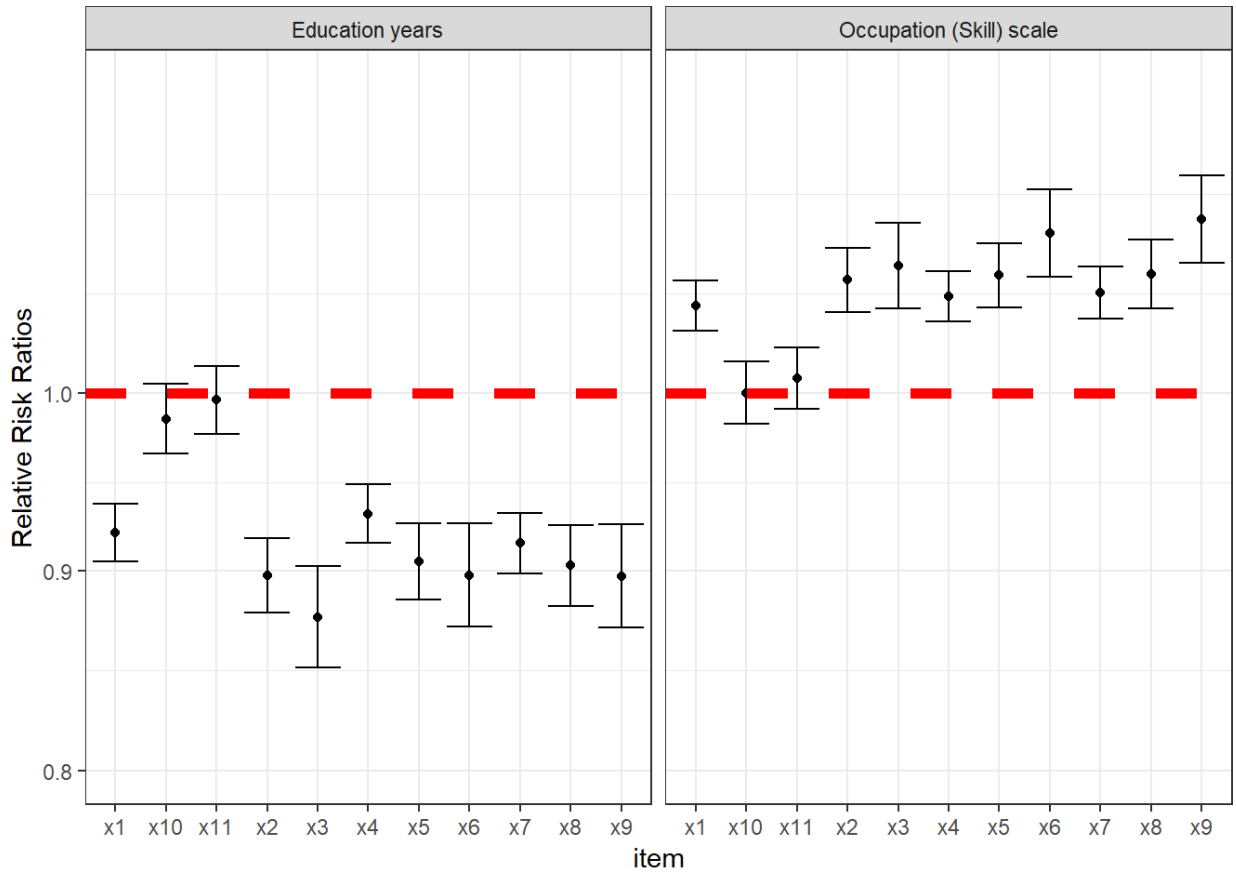

Figure 4.5: This plot shows the Relative Risk Ratios for each item using two validators (adjusted by the total household members)

```
## 22 0.9958716 0.9759022 1.0162522  x11          Education years
```

Once the models have been fitted, we could proceed to inspect the parameters. To inspect them we produce two plots shown in figure~4.5. The plot show the RRRs for both education and occupation adjusted by the household size. There is no evidence to support an association between items x10 and x11 and both predictors of poverty. In contrast, education and occupation predict an decrease and increase in the likelihood of being deprived of items x1-x9. On this basis we could conclude that our scale has criterion validity.

```
p<- ggplot(coefs, aes(x=item,y=X1)) + geom_point() +
  geom_errorbar(aes(ymin=X2, ymax=X3)) +
 theme_bw() + scale_y_continuous(trans = 'log10', limits = c(.8, 1.2))
p + facet_grid(. ~ var) + labs(y="Relative Risk Ratios") + geom_hline(yintercept=1, linetype="dashed",
                color = "red", size=2)
```

```
p<- ggplot(coefs, aes(x=item,y=X1)) + geom_point() +
  geom_errorbar(aes(ymin=X2, ymax=X3)) +
 theme_bw() + scale_y_continuous(trans = 'log10', limits = c(.8, 1.2))

jpeg("val_rrrs.jpg", units="cm", width=10, height=10, res=300)
p + facet_grid(. ~ var) + labs(y="Relative Risk Ratios") + geom_hline(yintercept=1, linetype="dashed",
                color = "red", size=2)
dev.off()
```

4.4.2 Construct Validity

Validity now is seen under a unified approach that looks at different aspects of the extent to which our scale can be interpreted as it is supposed to- a measure of poverty. Predictive validity might be a useful way to check the predictive validity at item-level. However, such kind of validation tells nothing about the structure of the measure. In section @ref{} we mention that modern poverty research should walk toward the specification of measurement models so that researchers make their assumptions better. We have mention that our scale is a higher-order scale with three dimensions, each one measured by three items. Construct validity concerns with the assessment of the structure of our scale. We will address several hypothesis about our scale:

- Are three dimensions a sensible way to arrange our indicators?
- Is a higher order factor present in our scale?
- Is the contribution to the explanation of the variance of each item equal or unequal?

We will focus on the first two question for now. To assess the validity of our measure we will use CFA to assess whether our measurement model is an adequate representation of poverty given these data. A CFA explicitly asks the question about the capacity of a model to reproduce the observed data. The first step, thus, consists in specifying our model. We have done already this in section @ref{} when we estimated the reliability statistics ω and ω_h. We will fit again the model using the `lavaan` R-package and Mplus. We will start with `lavaan` by specifying the MD_model. As can be appreciated we are assuming three factors (f1 to f3) and a higher order factor h. We are also stating that the indicators are manifest of one factor, i.e. we do not see x1 in f2 or f3. Then we can simply use the `sem()` function and tell that our items are categorical. We will store the output in the fit object.

```
MD_model <- ' f1  =~ x1 + x2 + x3
              f2 =~ x4 + x5 + x6
              f3   =~ x7 + x8 + x9
               h =~ f1 + f2 + f3
'

fit <- sem(MD_model,
           data = Rel_MD_1,ordered=c("x1","x2","x3","x4","x5",
                                     "x6","x7","x8","x9"))
```

Once the model has been estimated, we can request the global statistics of fit of our model saved in the fit object. To extract the statistics we will use the function `fitmeasures()`. We will request the χ^2 test (absolute fit), the CFI and TLI values and RMSEA (relative fit). The p-value of the χ^2 test suggest that we reject the hypothesis that the model does not reproduces the observed data. That means that dimensions, classification of the indicators and the presence of the higher order factor do a good job in representing the structure of the data. CFI, TLI and RMSEA point in the same direction.

```
chisq<-fitmeasures(fit,fit.measures = c("chisq","df","pvalue"))
relfit<-fitmeasures(fit,fit.measures = c("tli","cfi"))
rmsea<-fitmeasures(fit,fit.measures = c("rmsea", "rmsea.ci.lower",
                                        "rmsea.ci.upper", "rmsea.pvalue"))
chisq
```

```
##  chisq     df pvalue
## 17.717 24.000  0.817
```

```
relfit
```

```
## tli cfi
##   1   1
```

```
rmsea
```

```
##          rmsea rmsea.ci.lower rmsea.ci.upper   rmsea.pvalue
```

```
##          0.000          0.000          0.007          1.000
```

We can fit the same model in Mplus using the following code. We estimate the same model: three dimensions, one higher-order factor and each dimension with three exclusive indicators. We store the model specification in the test object and then we use the function `mplusModeler()` to pass (rel_CFA_1.inp) and fit the model on Mplus. The results of this operation are saved on the `res` object.

```
test <- mplusObject(
TITLE = "Higher order CFA;",
   VARIABLE = "
     NAMES = x1-x11 resources educ_yr occupation hh_size class;
     CATEGORICAL = x1-x9;
     USEVARIABLES = x1-x9;",
   ANALYSIS = "ESTIMATOR = WLSMV;
               PROCESS = 4",

MODEL = "f1 by x1-x3;
  f2 by x4-x6;
  f3 by x7-x9;
  h by f1 f2 f3;",

OUTPUT = "STD stdyx;")

res <- mplusModeler(test, modelout = "rel_CFA_1.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="Rel_MD_data_1_1.dat", run = 1L)
```

```
##
## Running model: rel_CFA_1.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "rel_CFA_1.inp"
## Reading model:  rel_CFA_1.out
```

Once the model has been estimated we can request the global statistics of fit using the following piece of code. We observed that the estimates match the `lavaan()` figures. The model reproduces the observed data.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate)
fitstats
```

```
##    TLI    CFI  Chisq  RMSEA
## 1.0000 1.0000 0.1576 0.0080
```

4.4.3 A joint assessment: Criterion and construct validity

Ideally, we would like to move toward a unified validation of scales. This involves examining both criterion and construct validity in the same model. Previously, we discussed that our full model looks like figure 4.3. This is called a MIMIC model. This moves us from the world of CFA into Structural Equation Modelling (SEM) but still the focus is on measurement and not so much on explanation. Again we will use `lavaan()` and Mplus to fit the model. In `lavaan()` we just need to create a new model that includes a new path. We would like to assess whether the higher-order factor (h) is associated with resources, adjusting by the total of household members. This can be simply achieved by adding a new line with a regression of h on the variables resources and hh_members. We fit and save the model in the fit object.

```
MD_model <- ' f1  =~ x1 + x2 + x3
              f2 =~ x4 + x5 + x6
              f3   =~ x7 + x8 + x9
               h =~ f1 + f2 + f3
               h ~ resources + hh_members
'

fit <- sem(MD_model,
           data = Rel_MD_1,ordered=c("x1","x2","x3","x4","x5",
                                     "x6","x7","x8","x9"))
```

Construct validity is assessed on the same terms. We will look at the overall fit of our model, which now know includes a new path, using the same statitsics: χ^2, CLI, TLI and RMSEA. We find that our measurement model still holds.

```
chisq<-fitmeasures(fit,fit.measures = c("chisq","df","pvalue"))
relfit<-fitmeasures(fit,fit.measures = c("tli","cfi"))
rmsea<-fitmeasures(fit,fit.measures = c("rmsea", "rmsea.ci.lower",
                                        "rmsea.ci.upper", "rmsea.pvalue"))
chisq
```

```
##  chisq     df pvalue
## 26.066 40.000  0.956
```

```
relfit
```

```
## tli cfi
##   1   1
```

```
rmsea
```

```
##          rmsea rmsea.ci.lower rmsea.ci.upper   rmsea.pvalue
##              0              0              0              1
```

Now we can check criterion validity by looking at the parameters of the regression part of our model. To extract the values of the parameters we will use the function `parameterEstimates()`, which is applied to the object fit. This is save in the slope object, which has all the estimated parameters in our model. For simplicity we will only show the slope h on resources by selecting the appropriate row. We observe that indeed there is a relationship between the factor and our parameters. What is the meaning of the reported value? The factor scores are presumed to follow a normal distribution. The higher the values of the factor, the higher the severity of poverty and vice versa. Therefore, we see that higher resources predict a decrease in the factor score, which is the expected behaviour in our measurement model.

```
slope<-as.data.frame(parameterEstimates(fit))
slope[13,]
```

```
##    lhs op       rhs         est           se        z pvalue    ci.lower
## 13   h  ~ resources -0.02989634 0.0009216571 -32.4376      0 -0.03170276
##       ci.upper
## 13 -0.02808993
```

We can estimate the same Model in Mplus as follows. All we have to do is to add a new path "h on resources and hh_members". The rest of the script is similar to the previous CFA model. We will create the following Mplus syntax: val_sem_1.inp. We will run the model using the `mplusModeler()` function and ask R to run the model in Mplus and save everything in the `res` object.

```
test <- mplusObject(
TITLE = "
HIgher order MIMIC;",
```

```
VARIABLE="

     NAMES = x1-x9 resources educ_yr occupation hh_members class;
     CATEGORICAL = x1-x9;
     USEVARIABLES = x1-x9 resources hh_members;",
ANALYSIS="
ESTIMATOR = WLSMV;
               PROCESS = 4;",
MODEL= "
f1 by x1-x3;
  f2 by x4-x6;
  f3 by x7-x9;

h by f1 f2 f3;

h on resources hh_members;",

OUTPUT=
"STD stdyx;")

res <- mplusModeler(test, modelout = "val_sem_1.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="Rel_MD_data_1_1.dat", run = 1L)
```

```
## Wrote model to: val_sem_1.inp
```

```
## Wrote data to: Rel_MD_data_1_1.dat
```

```
## No action taken as writeData = 'never'
```

We extract the parameters of the MIMIC model using `{fitmeasures()` and then we check the estimates. We confirm that our estimates reproduce the `lavaan()` output.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate)
fitstats
```

```
##    TLI    CFI  Chisq  RMSEA
## 0.9990 1.0000 0.0603 0.0090
```

With some code we could request the estimate of the slope as we did with the `lavaan()` model. However, we will stress the importance of visualising our measurement models by looking at the standardised parameters on a diagram of our model. Figure **??**(fig:valsem1) shows the standardised estimates of our model. We can see that resource predicts poverty -latent factor- (following Townsnend's theory representation in this case) and this constitutes a validation of our measure.

```
knitr::include_graphics("val_sem_1.png")
```

4.4.4 Real-data example

We will use the Mexican data (pobreza_14.dta) to illustrate how validity could be assessed using a MIMIC model. We had already created a *.dat file (Mex_pobreza_14.dat) with the variables we need for the analysis (Section). We can inspect the deprivation variables to familiarise ourselves with these data. The reduce model

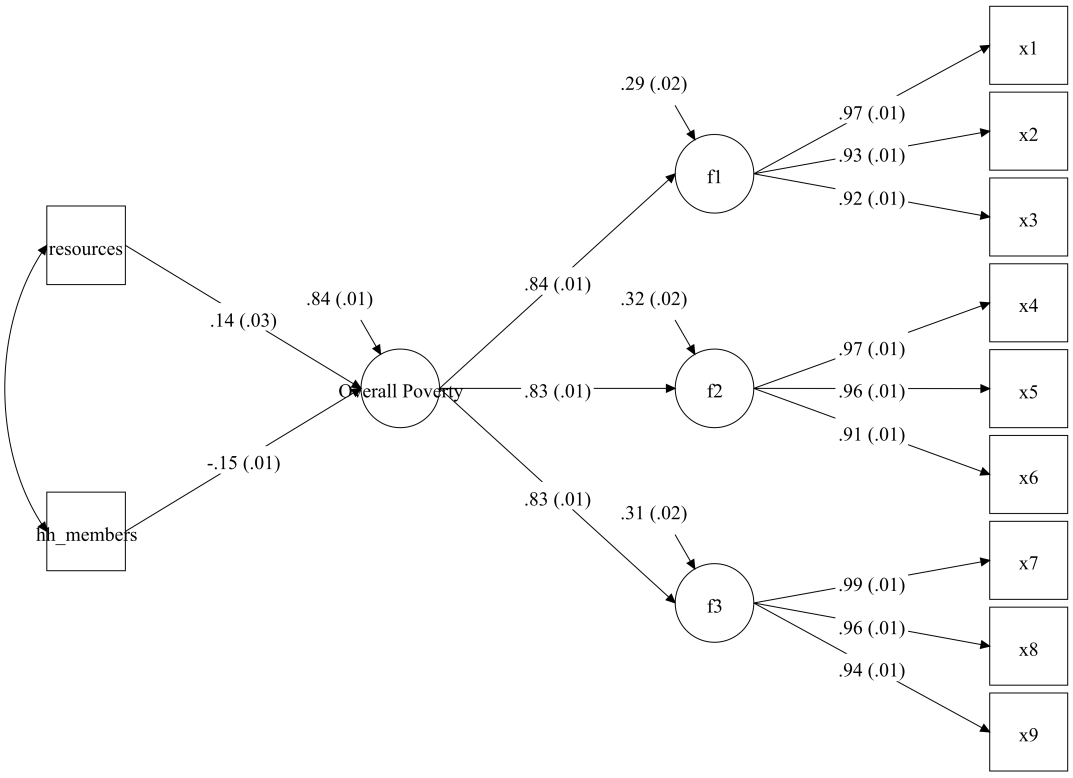

Figure 4.6: This is a MIMIC model were a higher-order factor model loads into three dimensions and there is one path to examine criterion validity (resources and hh members)

of the Mexican multidimensional measure comprises 14 variables classified in three dimensions: Housing, Essential services and food deprivation.

```
library(haven)
Mex_D<-read_dta("pobreza_14.dta")
head(Mex_D[31:34])
```

```
## # A tibble: 6 x 4
##            icv_muros          icv_techos         icv_pisos          icv_hac
##            <dbl+lbl>           <dbl+lbl>         <dbl+lbl>        <dbl+lbl>
## 1 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
## 2 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
## 3 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
## 4 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
## 5 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
## 6 0 [No presenta ca~ 0 [No presenta car~ 0 [No presenta c~ 0 [No presenta ~
```

```
head(Mex_D[36:39])
```

```
## # A tibble: 6 x 4
##             isb_agua           isb_dren          isb_luz         isb_combus
##            <dbl+lbl>          <dbl+lbl>        <dbl+lbl>          <dbl+lbl>
## 1 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
## 2 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
## 3 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
## 4 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
## 5 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
## 6 0 [No presenta ca~ 0 [No presenta ca~ 0 [No presenta ~ 0 [No presenta ca~
```

```
head(Mex_D[41:46])
```

```
## # A tibble: 6 x 6
##   ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad
##    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1      1      1      1      1      1      0
## 2      0      0      0      0      0      0
## 3      1      0      0      0      0      0
## 4      0      0      0      0      0      0
## 5      0      0      0      0      0      0
## 6      1      1      1      0      1      0
```

For the validity analysis we will fit the same higher-order CFA model with the three dimensions (essential services, housing quality and food deprivation). We will use two validators: Education attainment of the household head and an index of assets (fridge, tv, washing machine, computer and internet). The estimation of the parameters of both validators will be adjusted by rural v urban areas (rururb) and household size (tot_integ).

We will fit the model on Mplus. We will add the four new variables to the USEVARIABLES list and then include four new paths to the CFA model. This is achieved by including "h on rururb tot_integ durables educ_hh;" in the script (val_CFA_mex.inp). Again, we will save this in the test object and we will run the model from R using the `mplusModeler()` function. Bear in mind that the model will take some seconds to run.

```
test <- mplusObject(
TITLE = "Validity Mexico CFA model;",
   VARIABLE = "
     NAMES = proyecto folioviv foliohog icv_muros icv_techos
             icv_pisos icv_hac isb_agua isb_dren isb_luz isb_combus
```

```
             ic_sbv ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad
             ia_7men ia_8men ia_9men ia_10men ia_11men ia_12men
             tv_dep radio_dep fridge_dep
             washingmach_dep compu_dep inter_dep psu weight
             rururb tot_integ durables educ_hh;
MISSING=.;
     CATEGORICAL = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
     USEVARIABLES = icv_muros icv_techos icv_pisos icv_hac isb_agua
                   isb_dren isb_luz isb_combus  ia_1ad
                   ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad
                   rururb tot_integ durables educ_hh;

WEIGHT=weight;
cluster = psu;",

   ANALYSIS = "TYPE = complex;

ESTIMATOR = wlsmv;
PROCESS = 4;",

MODEL = "f1 by icv_muros icv_techos icv_pisos icv_hac;
  f2 by isb_agua
        isb_dren isb_luz isb_combus;
  f3 by ia_1ad ia_2ad ia_3ad ia_4ad ia_5ad ia_6ad;
  h  by f1 f2 f3;
  h on durables educ_hh rururb tot_integ;",

OUTPUT = "std stdyx;")

res<-mplusModeler(test, modelout = "val_CFA_mex.inp",
                    writeData = "never", hashfilename = FALSE,
                    dataout="Mex_pobreza_14.dat", run = 1L)
```

```
## Wrote model to: val_CFA_mex.inp

## Wrote data to: Mex_pobreza_14.dat

## No action taken as writeData = 'never'

##
## Running model: val_CFA_mex.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "val_CFA_mex.inp"
## Reading model:  val_CFA_mex.out
```

Once the model has been fitted we can examine construct validity by assessing whether our model holds after adding the predictors. We will save the statistics of fit in the `fitstats()` object. We can see that the fit of this model is very good. We find that the model seems to be a valid representation of poverty for Mexico. That means that the dimensions and indicators are adequately classified and identified.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate)
fitstats
```

```
##   TLI   CFI Chisq RMSEA
## 0.987 0.989 0.000 0.019
```

On Mplus we can produce a diagram to display the estimated values of the parameters of our model. From left to right, we appreciate the standardised parameters of the validators and the adjustment variables. We see the four have the expected signs. The higher the education attainment, the lower the factor scores (higher severity). The asset index shows a similar behaviour, having more durables in the household is associated with lower factor scores. Both rurality and the household size increase the factor scores. Then we appreciate that the standardised factor loadings are high ($> .5$).

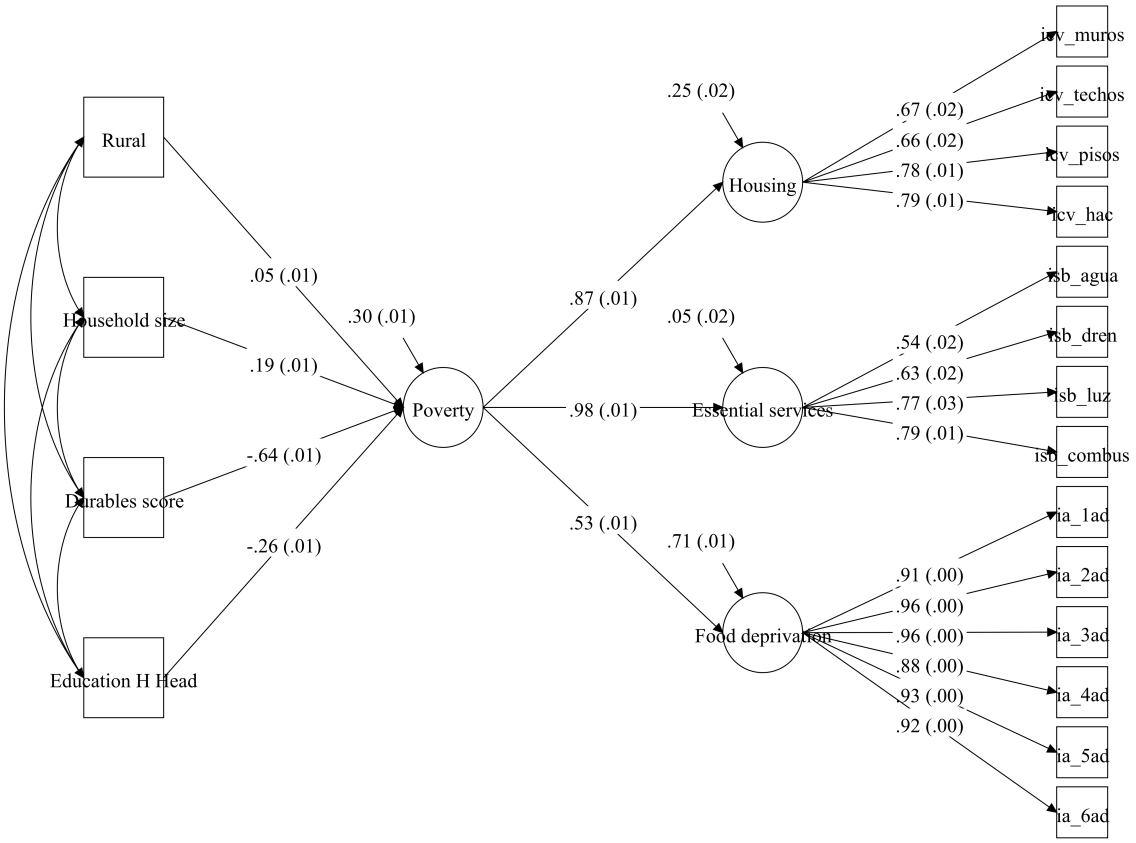

Figure 4.7: This is a MIMIC model of a reduced version of the multidimensional Mexican measure. The model shows that poverty is associated by possession of different goods and education attainment of the household head, adjusted by rurality and household size. Standardised coefficients (Standard error within brackets)

Chapter 5

Comparability in poverty measurement

Abstract

This chapter discusses the problem of comparability in poverty measurement and frames this challenge using the principle of measurement invariance. The chapter provides an intuitive explanation of measurement invariance. A connection between measurement invariance, validity and reliability is provided. Then the chapter uses simulated data to illustrate how measurement invariance works and how can it be analysed in Mplus. The chapter also provides a real-data example.

5.1 Measurement invariance

Making comparisons of poverty across time or units (countries, regions, population groups) is one of the chief goals in poverty research. Ideally, changes in poverty from a given year to another must reflect effective changes in living standards. The same must hold when comparing estimates across groups where poverty must be relatively higher or lower given differences in living standards across the units of interest. Nonetheless, survey data is subject to different kinds of amendments overtime. There are changes to the questionnaire that aim to update poverty indices, researchers also include or exclude different indicators overtime and across groups, data collection modes and sampling frameworks might differ. These modifications are likely to affect the comparability of poverty estimates overtime or across countries using similar data.

In poverty measurement the literature often proposes that if one has the same indicators, poverty is being measured on equivalent terms. MI transforms this proposition into an assumption as there is no guarantee that using the same indicators measure poverty on the same terms. There are several survey and non-survey aspects affecting the full comparability of poverty indices: sampling framework, data-collection mode (face to face, computer-based, telephone, etc.), different survey questionnaires, changes in living standards that result between-unit reliability and validity discrepancies (i.e. some items or dimensions might be highly reliable to measure poverty in one group but not in other). The effect of these issues could be so high upon comparability than just using the same indicators to contrast the severity and prevalence of poverty across groups or time is unlikely to be enough.

The concern with the quantifying the effect of the different sources of incompatibility and disparity of different indices pushed measurement theory to develop a framework to conceptualise and then empirically assess comparability across measures. The rationale of factor analysis - the existent of a measurement model for a population- was extended to assess the extent to which different indices are comparable or not. This has resulted in the development of the concept of measurement invariance (MI) which defines comparability in terms of the extent to which a factor model holds for different populations [Meredith (1993);Meredith & Teresi (2006);Schoot, Lugtig, & Hox (2012);Lubke, Dolan, Kelderman, & Mellenbergh (2003);Byrne, Shavelson, &

Muthén (1989)}. That means that the structure of a poverty index (dimensions and items), the relationship between the indicators and the latent construct and the error term are similar across periods or groups. That is, having two poverty indices with the same indicators is not a sufficient to make meaningful comparisons of the prevalence and severity of poverty across groups. Therefore, Measurement Invariance (MI) is a necessary condition for making comparisons of subjects using a given index. Formally, MI can be defined as the capacity of an index to measure equivalently across two or more groups or periods.

MI implies that deprivation should change equally across two groups after the level of poverty of two groups changes in the same order of magnitude. In contrast, when MI is violated, it means that one group is being unfairly compared against another because there is another phenomenon causing the change in deprivation. This is a very undesirable feature of a poverty index because it would mean that the differences in prevalence and severity are due to different sources that are not related with poverty. Researchers cannot conclude that poverty is higher/lower given that the discrepancies are explained by sampling differences, data collection modes, dissimilar questionnaires, acute between-group differences reliability and validity.

MI is an ideal preposition and it could be violated in different ways. There are diverse aspects of MI that can be assessed and translated into the following standards (Meredith, 1993).

- Strict MI: This is the ideal level of MI as the structure, the relationship of the items with the latent variable, the indicator means and the residuals are equivalent across groups/periods.
- Strong MI: The structure, the relationship of the items with the latent variable and the indicator means are invariant across groups/periods.
- Weak MI: The structure and the relationship of the items with the construct are equivalent.
- Configural MI: Only the structure is the same across groups/periods.

5.2 Introduction to key aspects of measurement invariance

Measurement invariance is, therefore, about the similarity of the different parameters of a latent variable model across different groups. To introduce the notion of MI, simulated data was generated for 20 groups. In the simplest case, a unidimensional model is proposed where poverty is measured using 15 binary indicators. To make clearer what could happen when MI is violated, the parameters of five indicators where changed for half of the groups. This poses a situation where the poverty indices across 20 groups (e.g. countries) have the same structure -Configural Invariance- but there are substantive differences in the way in which some indicators capture poverty across units.

A two Item Response Theory (IRT) model was fitted to the simulated data to estimate both discrimination (a) and severity (b) parameters for each group. The models were fitted on Mplus 7.2 using the following code and the R-package MplusAutomation (Hallquist & Wiley, 2018).

```
[[init]]
iterators = i j;
i = 1:10;
j = 1:2;
filename = IRT_[[j]]_[[i]].inp;

outputDirectory = "C:../PM Book";

[[/init]]
DATA : FILE = UD_data_[[j]]_[[i]].dat;

VARIABLE : NAMES=V1-V15;
           USEVARIABLES=V1-V15;
           CATEGORICAL = V1-V15;

MODEL: f by V1-V15*;
       f@1;
```

```
## Running the IRT using the Mplusautomation package ##

#createModels("IRT_models_MI_Section.txt")
#runModels(filefilter = "MI_IRT_")

#Importing the *.out from mplus into R#

irt_MI<-readModels(filefilter ="mi_irt_")

#Putting both parameters into a list#

irt_MI<-lapply(irt_MI, function(x) {
  x<-x$parameters$irt.parameterization
  x<-x[1:30,]
  x<-data.frame(a=x$est[1:15],b=x$est[16:30])
  x
}
)
```

We inspect the first object in list `irt\_MI` and check the values of both discrimination (a) and severity (b) for the 15 items. The items have increasing discrimination values.

```
irt_MI[[1]]
```

```
##        a     b
## 1  0.667 2.905
## 2  0.861 2.292
## 3  0.791 2.510
## 4  0.911 2.340
## 5  0.981 1.586
## 6  1.167 1.126
## 7  1.214 0.871
## 8  1.301 0.651
## 9  1.285 0.457
## 10 1.363 0.724
## 11 1.405 0.677
## 12 1.437 0.698
## 13 1.473 0.427
## 14 1.564 0.287
## 15 1.699 0.197
```

The list has 20 objects it is possible to create a data frame using few lines of code so that we can plot and inspect how the estimates vary across groups.

```
#Load this packages if necessary
library(ggplot2)
library(reshape2)

#Creating a data frame to plot the parameters by group
irt_MI<-as.data.frame(irt_MI)
irt_MI$var_id<-1:15
```

The values of both parameters for group 1 and 10 can be inspected using the following code. It is clear that there are very little deviation from one group to another. This is an indication than strong MI might hold between these two groups as both loadings (discrimination) and thresholds (severity) are similar between group 1 and 10. However, the interest is to assess whether the parameters change dramatically across groups.

```
irt_MI[1:15,1:4]
```

```
##    mi_irt_1_1.out.a mi_irt_1_1.out.b mi_irt_1_10.out.a mi_irt_1_10.out.b
## 1             0.667            2.905             0.676             2.875
## 2             0.861            2.292             0.738             2.470
## 3             0.791            2.510             0.799             2.566
## 4             0.911            2.340             0.905             2.353
## 5             0.981            1.586             0.982             1.590
## 6             1.167            1.126             1.101             1.152
## 7             1.214            0.871             1.099             0.906
## 8             1.301            0.651             1.254             0.646
## 9             1.285            0.457             1.268             0.426
## 10            1.363            0.724             1.361             0.692
## 11            1.405            0.677             1.373             0.658
## 12            1.437            0.698             1.467             0.649
## 13            1.473            0.427             1.449             0.356
## 14            1.564            0.287             1.610             0.253
## 15            1.699            0.197             1.560             0.169
```

Then we can rearrange the data to produce the plots to make a visual inspection of the parameters.

```
irt_MI<-melt(irt_MI, id="var_id")
irt_MI$variable<-sub('.*(?=.$)', '', irt_MI$variable, perl=T)
irt_MI$data<-rep(1:20,each=30)

irt_MIa_mi<-subset(irt_MI,irt_MI$variable=="a" & irt_MI$var_id>=5)
irt_MIa_nonmi<-subset(irt_MI,irt_MI$variable=="a" & irt_MI$var_id<5)

irt_MIb_mi<-subset(irt_MI,irt_MI$variable=="b" & irt_MI$var_id>=5)
irt_MIb_nonmi<-subset(irt_MI,irt_MI$variable=="b" & irt_MI$var_id<5)
```

Figure 5.1 plots the discrimination parameters of each group that are likely -we say likely because below MI is formally tested- to be invariant across the 20 groups. There are small fluctuation from one group to another, indicating that changes in poverty produce similar changes in deprivation of the item in question across groups. As discussed in Chapter 4 (Reliability), this parameter can be used to assess monotonicity. Figure 5.2 plots the discrimination parameters for the first five items (1 to 5). For the first 10 groups the fluctuation is small. However, we can appreciate that for the other 10 groups, the discrimination parameters are very different. This is an indication that MI might not hold between two clusters of groups (1-10 and 11-20). The discrimination values in plot 5.2 suggest that a change in poverty result in less dramatic changes in deprivation for the first ten groups. If these five indicators were exclusively used to measure poverty, some groups will be highly disfavoured relative to the others as there is something else (not only poverty) causing changes in observed deprivation. In the next section the discussion is enriched by using a real data example.

```
ggplot(irt_MIa_mi,aes(x=data,y=value,group=var_id)) + geom_point()  +
  geom_line(aes(linetype=as.factor(var_id))) +
  xlab("Groups 1-20") + ylab("Discrimination parameter. IRT scale") +
  labs(linetype='Indicator id') +
  scale_y_continuous( limits = c(.5,2), expand = c(0,0), breaks = seq(.5, 2, .25) ) +
  theme_bw()
```

```
ggplot(irt_MIa_nonmi,aes(x=data,y=value,group=var_id)) + geom_point()  +
  geom_line(aes(linetype=as.factor(var_id))) +
  xlab("Groups 1-20") + ylab("Discrimination parameter. IRT scale") +
  labs(linetype='Indicator id') +
  scale_y_continuous( limits = c(.5,2), expand = c(0,0), breaks = seq(.5, 2, .25) ) +
  theme_bw()
```

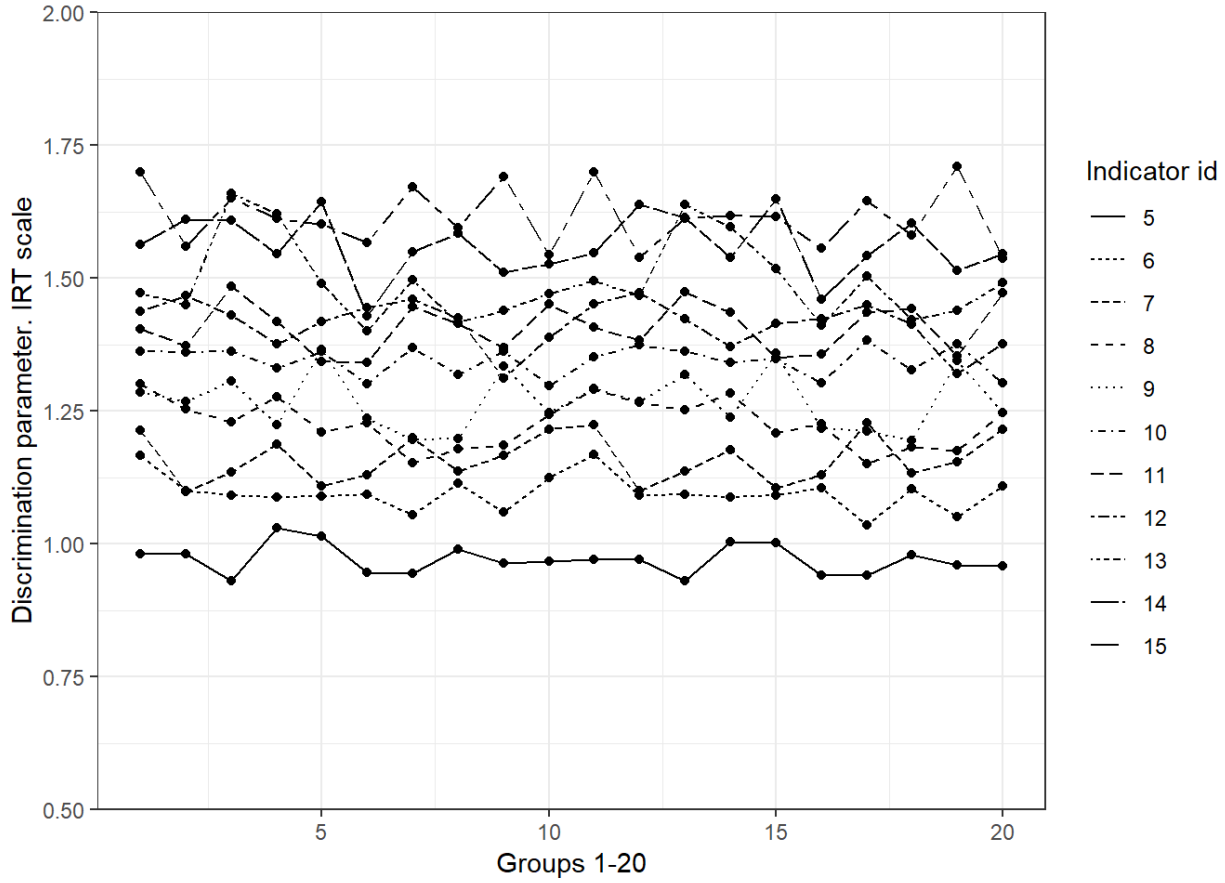

Figure 5.1: Discrimination parameters that seem to fulfil MI. Simulated data. We see very little fluctuation from one group to another

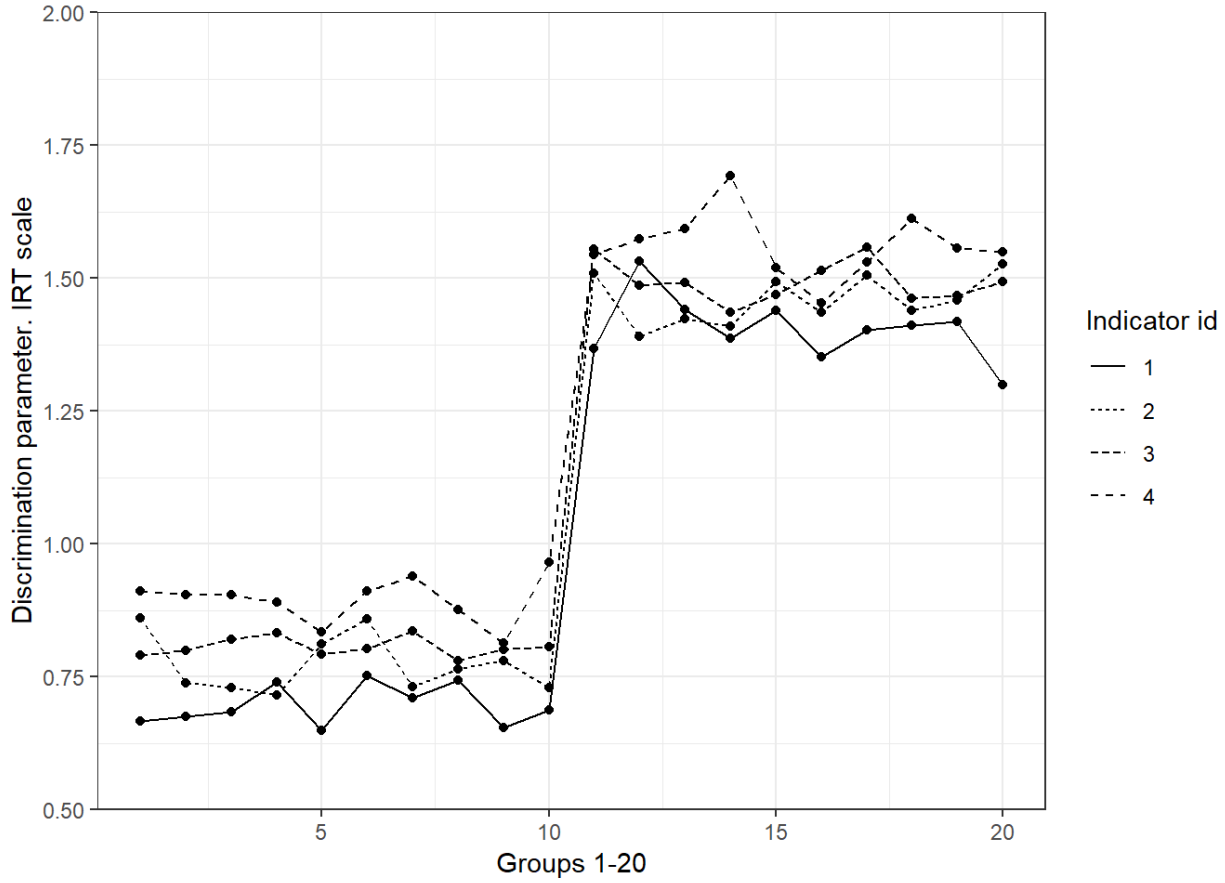

Figure 5.2: Discrimination parameters that do not seem to fulfil MI. Simulated data. We see a lot of fluctuation from groups 11-20 relative to 1-10

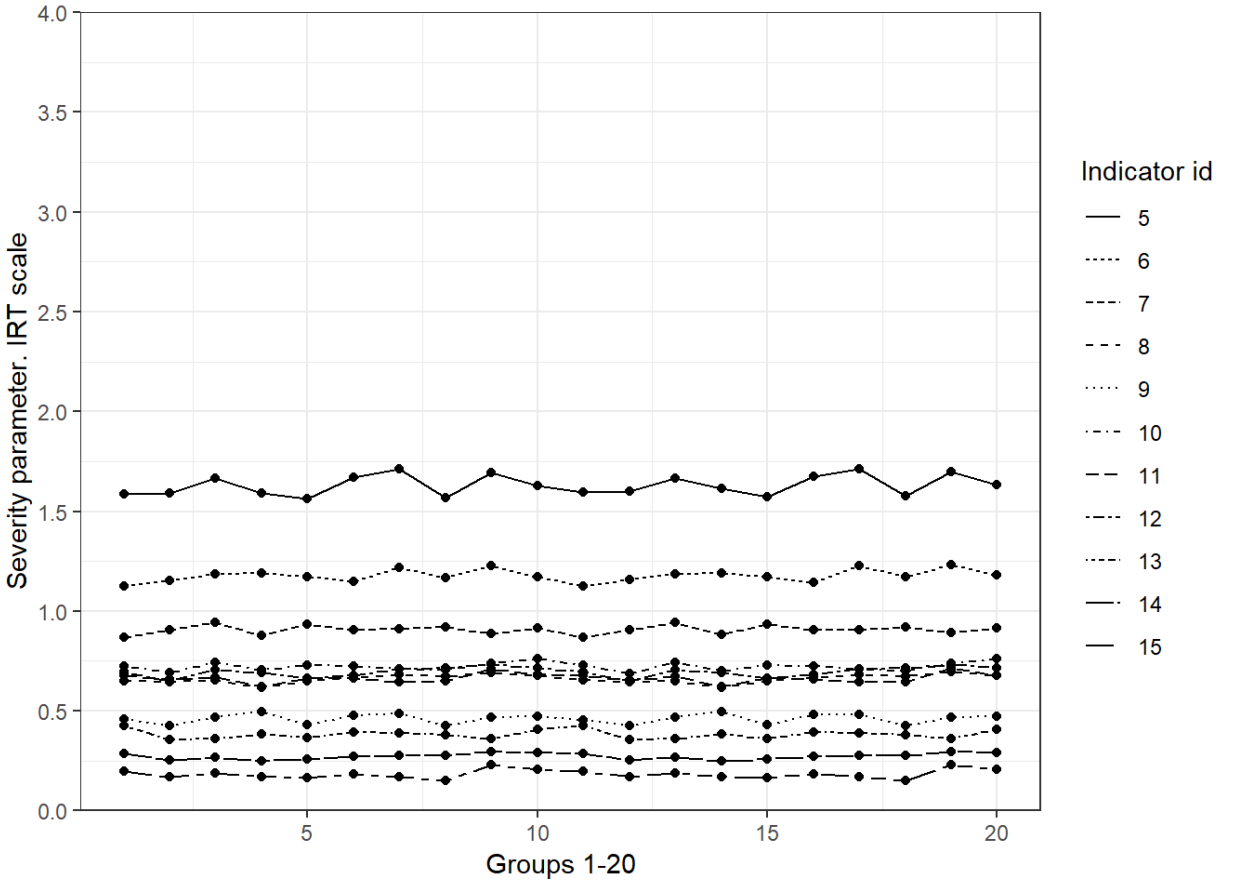

Figure 5.3: Severity parameters that seem to fulfil MI. Simulated data. We see very little fluctuation from one group to another

Figures 5.3 and 5.4 plot the values of the severity parameter for the items that are likely to be invariant across all groups and for the items that seem to violate MI, respectively. The severity parameters range between 0 and 3 standard deviations. As discussed in Chapter 4, this is the expected behaviour when looking at low standard of living. The items in plot 5.3 are very stable across groups. In contrast the severity parameters change a lot for groups 11 to 20 in comparison with the first ten groups (Figure 5.4). It seems, therefore, that these items are non-invariant, at least, between these two clusters. The severity parameter is tied with the intercept of a factor model (See Chapter 4). It indicates the mean value of deprivation for each item given the latent value of poverty. Different means are a violation of strong MI, and therefore is undesirable. In poverty research indicates the case where an indicator is a more/less severe manifestation of poverty when comparing two or more groups. In other words, it is an indication of when an indicator of a given society might be too severe to measure poverty in another. This is explained further using the real data example.

```
ggplot(irt_MIb_mi,aes(x=data,y=value,group=var_id)) + geom_point()  +
  geom_line(aes(linetype=as.factor(var_id))) +
  xlab("Groups 1-20") + ylab("Severity parameter. IRT scale") +
  labs(linetype='Indicator id') +
  scale_y_continuous( limits = c(0,4), expand = c(0,0), breaks = seq(0, 4, .5) ) +
  theme_bw()
```

```
ggplot(irt_MIb_nonmi,aes(x=data,y=value,group=var_id)) + geom_point()  +
  geom_line(aes(linetype=as.factor(var_id))) +
  xlab("Groups 1-20") + ylab("Severity parameter. IRT scale") +
```

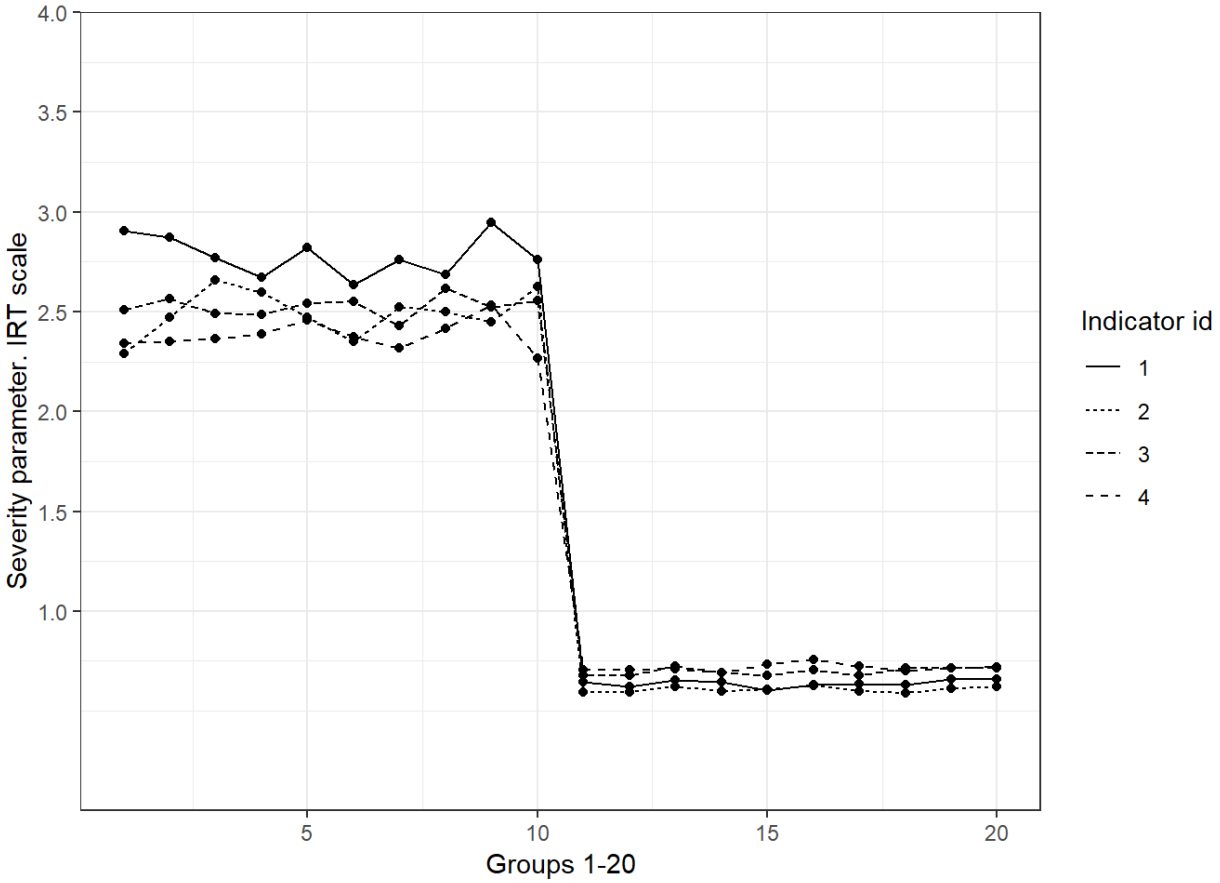

Figure 5.4: Severity parameters that do not seem to fulfil MI. Simulated data. We see a lot of fluctuation from groups 11-20 relative to 1-10

```
  labs(linetype='Indicator id') +
  scale_y_continuous( limits = c(0,4), expand = c(0,0), breaks = seq(1, 4, .5) ) +
  theme_bw()
```

5.3 Methods for the assessment of Measurement Invariance

Measurement Invariance is formally examined by comparing the extent two which the parameters of a measurement model are similar between two or more groups/periods. The measurement model can be a unidimensional or a multidimensional factor model. That is, it can be seen as a formal assessment of the visual inspection shown in the previous section. The literature proposes two main methods to analyse MI: Multiple Group Factor Analysis (MGFA) and the Alignment Method (AM). The rationale behind both methods is the same, start from a measurement model with some free parameters (configural model, for example) and then move toward a model with fixed parameters (strong MI where loadings and thresholds are fixed across groups). The different models are contrasted using both absolute statistics of fit such as Chi-Square and relative statistics of fit like RMSEA and TLI, CFI. Based on these statistics researchers conclude whether the same model holds for two different groups.

```
ud_1<-read.table("UD_data_2_1.dat")
ud_1$data<-2
ud_2<-read.table("UD_data_1_1.dat")
ud_2$data<-1
```

```
ud_1and2<-rbind(ud_1,ud_2)
head(ud_1and2)
```

```
##   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 data
## 1  0  1  0  1  1  0  0  1  0   0   0   0   1   1   1   2    2
## 2  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   2    2
## 3  1  1  1  1  1  1  0  0  1   1   1   0   1   1   1   2    2
## 4  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   1    2
## 5  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   1    2
## 6  0  0  0  1  0  0  0  0  0   0   0   0   0   1   1   2    2
```

```
write.table(ud_1and2, file="ud_1and2.dat", col.names = F)
```

5.3.1 Multiple Group Factor Analysis

Multiple Group Factor Analysis had been the classic method to assess MI. To illustrate how it works on Mplus ("MI_analysis_du_configural1and2.inp"), data for group 1 and group 11 ("ud_1and2.dat") will be used to go through the different steps involved in testing MI. The first step consist in testing the hypothesis that a configural model holds between these two groups, i.e. whether the same unidimensional model holds leaving the loadings and thresholds free across groups. The syntax is displayed below. The syntax is very similar to the one used to fit a unidimensional Confirmatory Factor Model (CFA) following a two-parameter IRT model. One needs to tell Mplus the groups in question (GROUPING) and save the results of the Chi-square test to contrast out model with the models with fixed loadings and thresholds. It is also very important to request the Modification Indices, which tell which parameters are the sources of model miss fit and therefore the cause of violating more strict forms of MI. In this example, the mean of the latent variable is fixed to one given that those value were used to simulate these data.

```
test <- mplusObject(
TITLE = "Multiple Group Analysis;",
VARIABLE = "
Name = id V1-V16 data;
Missing are all (-9999) ;
usevariables = V1-V15;
categorical = V1-V15;
GROUPING = data (1=ONE 2=TWO);",

ANALYSIS="ESTIMATOR IS WLSMV;
PARAMETERIZATION=THETA;",

SAVEDATA= "DIFFTEST=Configural.dat;",

OUTPUT = "STDYX MODINDICES (3.84);",

MODEL = "

! Factor loadings all estimated
F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*] (T1-T15);

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
```

```
[f@0]; f@1;

!!! CONFIGURAL MODEL FOR ALTERNATIVE GROUP
MODEL TWO:  ! Factor loadings all estimated

F BY V1-V15*;

! Item thresholds all free
[v1$1-v15$1*];

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f@1;")

res <- mplusModeler(test, modelout = "MI_analysis_du_configural1and2.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="ud_1and2.dat", run = 1L)
```

```
##
## Running model: MI_analysis_du_configural1and2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "MI_analysis_du_configural1and2.inp
## Reading model:  MI_analysis_du_configural1and2.out
```

The fit of the model ("MI_analysis_du_configural1and2.inp") is shown below. The fit of the model is very good. Under the Chi-Square test the model is not rejected and the relative statistics of fit point in the same direction. However, this only a reference that allow us to inspect metric (weak) MI. Should the configural model is rejected, then comparing both poverty indices very likely lead to incorrect conclusion about both prevalence and severity of poverty.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate)
fitstats
```

```
##    TLI    CFI  Chisq  RMSEA
## 1.0000 1.0000 0.0133 0.0070
```

The Mplus input of the metric MI model, free threshold but fixed loadings, is displayed below ("MI_metricanalysis_du_1and2.inp"). To compare the fit of the configural model against the metric model it is necessary to call for the information stored in "DIFFTEST=Configural.dat;". The main difference between these INPUT INSTRUCTIONS and the configural model is that the loadings of the 15 items are fixed -equal across groups-. Therefore, the model will fit such assumption to the data and see whether it holds or not.

```
test <- mplusObject(
TITLE = "Multiple Group Analysis (Metric);",
VARIABLE = "
Name = id V1-V16 data;
Missing are all (-9999) ;
usevariables = V1-V15;
categorical = V1-V15;
GROUPING = data (1=ONE 2=TWO);",
```

```
ANALYSIS="ESTIMATOR IS WLSMV;
PARAMETERIZATION=THETA;
DIFFTEST= Configural.dat;",

SAVEDATA= "DIFFTEST=MetricA.dat;",

OUTPUT = "STDYX MODINDICES (3.84);",

MODEL = "

! Factor loadings all estimated
 F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*] (T1-T15);

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f@1;

!!! CONFIGURAL MODEL FOR ALTERNATIVE GROUP
MODEL TWO:  ! Factor loadings all estimated

 F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*];

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f*;")

res <- mplusModeler(test, modelout = "MI_metric_analysis_du_configural1and2.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="ud_1and2.dat", run = 1L)
```

```
##
## Running model: MI_metric_analysis_du_configural1and2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "MI_metric_analysis_du_configural1a
## Reading model:  MI_metric_analysis_du_configural1and2.out
```

The Chi-Square Test for Difference Testing in the output of the metric MI analysis leads to the rejection of the model ($Chi - Square < .05$). There is also a drop in the value of the relative statistics of fit TLI and CFI. These two suggest that although the model is relatively adequate, there is a loss after fixing the loadings across groups. This is expected given that the loadings of groups 1 and 11 are very different for some of the items.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
```

```
            RMSEA=res$results$summaries$RMSEA_Estimate,
            ChisqDIFF=res$results$summaries$ChiSqDiffTest_PValue)
fitstats
```

```
##       TLI       CFI     Chisq     RMSEA ChisqDIFF
##     0.992     0.993     0.000     0.031     0.000
```

The modification indices of the metric model provide information about the parameters that make the main contribution to the inadequate model fit. Figure 5.2 suggested that indicators 1-4 were very likely to be non-invariant. The modification indices confirm that indeed these items are the main sources of the discrepancy between the two groups. Without prior knowledge or a clear theory of why these four items have non-invariant loadings, the suggestion would be to let these four loadings free and re-assess metric invariance. Otherwise, the alternative would be dropping these four indicators and re-examine metric MI.

res*results*mod_indices[1:15,]

To assess whether partial metric MI holds, the loadings of the four items in question (1-4) were no longer fixed between groups. The syntax is omitted but the output ("mi_metricanalysis_b_du_1and2.out") is shown below. The Chi-Square Test for Difference Testing suggest that partial metric invariance holds. Indicating that it is feasible to proceed and conduct the examination of Strong MI (scalar MI).

```
test <- mplusObject(
TITLE = "Multiple Group Analysis (MetricB);",
VARIABLE = "
Name = id V1-V16 data;
Missing are all (-9999) ;
usevariables = V1-V15;
categorical = V1-V15;
GROUPING = data (1=ONE 2=TWO);",

ANALYSIS="ESTIMATOR IS WLSMV;
PARAMETERIZATION=THETA;
DIFFTEST= Configural.dat;",

SAVEDATA= "DIFFTEST=MetricB.dat;",

OUTPUT = "STDYX MODINDICES (3.84);",

MODEL = "

! Factor loadings all estimated
 F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*] (T1-T15);

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f@1;

!!! CONFIGURAL MODEL FOR ALTERNATIVE GROUP
MODEL TWO:

 F BY V1-V15* (L1a L2a L3a L4a L5-L15);
```

```
! Item thresholds all free
[v1$1-v15$1*];

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f*;")

res <- mplusModeler(test, modelout = "MI_metricanalysis_B_du_1and2.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="ud_1and2.dat", run = 1L)
```

```
##
## Running model: MI_metricanalysis_B_du_1and2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "MI_metricanalysis_B_du_1and2.inp"
## Reading model:  MI_metricanalysis_B_du_1and2.out
```

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate,
            ChisqDIFF=res$results$summaries$ChiSqDiffTest_PValue)
fitstats
```

```
##       TLI       CFI     Chisq     RMSEA ChisqDIFF
##    1.0000    1.0000    0.4639    0.0010    1.0000
```

In order to assess scalar invariance it is necessary to hold all the thresholds (intercepts) equal across groups.

```
test <- mplusObject(
TITLE = "Multiple Group Analysis (Scalar);",
VARIABLE = "
Name = id V1-V16 data;
Missing are all (-9999) ;
usevariables = V1-V15;
categorical = V1-V15;
GROUPING = data (1=ONE 2=TWO);",

ANALYSIS="ESTIMATOR IS WLSMV;
PARAMETERIZATION=THETA;
DIFFTEST= MetricB.dat;",

SAVEDATA= "DIFFTEST=Scalar.dat;",

OUTPUT = "STDYX MODINDICES (3.84);",

MODEL = "

! Factor loadings all estimated
 F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*] (T1-T15);
```

```
! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f@1;

!!! CONFIGURAL MODEL FOR ALTERNATIVE GROUP
MODEL TWO:

F BY V1-V15* (L1a L2a L3a L4a L5-L15);

! Item thresholds all fixed

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f*;")

res <- mplusModeler(test, modelout = "MI_scalaranalysis_du_1and2.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="ud_1and2.dat", run = 1L)
```

```
##
## Running model: MI_scalaranalysis_du_1and2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "MI_scalaranalysis_du_1and2.inp"
## Reading model:  MI_scalaranalysis_du_1and2.out
```

The output below ("mi_scalaranalysis_du_1and2.out") indicates that partial scalar invariance holds. It seems that for this data, it is just necessary to let the loadings of the first four items free to achieve partial scalar invariance. This suggests that after considering differences in slope the means between groups are equivalent. This does no necessarily mean that the scale is fully invariant, the effect of the first four items could lead to incorrect conclusions when comparing poverty levels between these two groups. Next section discusses the meaning of threshold non-invariance in the context of poverty research, as this topic connects with scale equating.

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate,
            ChisqDIFF=res$results$summaries$ChiSqDiffTest_PValue)
fitstats
```

```
##       TLI       CFI     Chisq     RMSEA ChisqDIFF
##     0.989     0.989     0.000     0.038     0.000
```

After fixing the thresholds of the first four items partial scalar invariance holds ($Chi - square > .05$). This would mean that in order to meet exact scalar invariance the four items in question must be dropped from the index. This is, of course, not ideal as most of the time poverty indices have few items to choose from.

```
test <- mplusObject(
TITLE = "Multiple Group Analysis (ScalarB);",
VARIABLE = "
Name = id V1-V16 data;
Missing are all (-9999) ;
```

```
usevariables = V1-V15;
categorical = V1-V15;
GROUPING = data (1=ONE 2=TWO);",

ANALYSIS="ESTIMATOR IS WLSMV;
PARAMETERIZATION=THETA;
DIFFTEST= MetricB.dat;",

SAVEDATA= "DIFFTEST=ScalarB.dat;",

OUTPUT = "STDYX MODINDICES (3.84);",

MODEL = "

! Factor loadings all estimated
 F BY V1-V15* (L1-L15);

! Item thresholds all free
[v1$1-v15$1*] (T1-T15);

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f@1;

!!! CONFIGURAL MODEL FOR WOMEN ALTERNATIVE GROUP
MODEL TWO:

F BY V1-V15* (L1a L2a L3a L4a L5-L15);

! Item thresholds four first fixed
[v1$1-v4$1*];

! Item residual variances all fixed=1
V1-V15@1;

! Factor mean=0 and variance=1 for identification
[f@0]; f*;")

res <- mplusModeler(test, modelout = "MI_scalaranalysis_B_du_1and2.inp",
                    writeData = "never",
                    hashfilename = FALSE,
                    dataout="ud_1and2.dat", run = 1L)
```

```
##
## Running model: MI_scalaranalysis_B_du_1and2.inp
## System command: C:\WINDOWS\system32\cmd.exe /c cd "." && "Mplus" "MI_scalaranalysis_B_du_1and2.inp"
## Reading model:  MI_scalaranalysis_B_du_1and2.out
```

```
fitstats<-c(TLI=res$results$summaries$TLI,
            CFI=res$results$summaries$CFI,
            Chisq=res$results$summaries$ChiSqM_PValue,
            RMSEA=res$results$summaries$RMSEA_Estimate,
```

```
            ChisqDIFF=res$results$summaries$ChiSqDiffTest_PValue)
fitstats
```

```
##       TLI       CFI     Chisq     RMSEA ChisqDIFF
##    1.0000    1.0000    0.7059    0.0000    0.9611
```

5.3.2 The alignment method

Multiple Group Factor Analysis has some disadvantages in real-data settings. In a real-data situation, researchers will have to fit a number of models to have an idea of the different sources of non-invariance. This is time-consuming and compromises the reproducibility of the findings given that another researcher might use different criteria to go through the modification indices. Another problem is that the MGFA is not feasible for many groups. Another disadvantage, when using Chi-Square, is that the models will be almost always rejected with large samples and researchers, therefore, have to work with rules of thumb to assess sufficient changes in TLI or CFI.

The alignment method has been put forward to overcome some of the drawbacks of MGFA. In particular, it simplifies the assessment of MI when having many groups. This method is under constant development but it aims to estimate means and variances of the latent factor conditional on a minimum level of MI. That means that it does not requires exact MI and aims at approximately MI. The alignment method therefore seeks an optional degree of measurement invariance given the data. The alignment method starts from the assumption that the configural model will be better that the fully scalar model. Once a configural model is fitted (M0), then the alignment method looks for a model that is equally as good as M0 but with some fixed parameters. By minimizing the loss function due to non-invariant parameters, the alignment method estimates factor means that are comparable conditional on the approximately invariant model.

```
temp = list.files(pattern="UD_data_.*.dat")
myfiles = lapply(temp, read.table)
myfiles<-myfiles[-11]
myfiles<-myfiles[-21]
data<-do.call(rbind,myfiles)
data$data<-rep(1:20,each=5000)
write.table(data,file="UD_MI_AM.dat", col.names = F)
```

To introduce the key aspects of the alignment method, this section relies on the simulated data generated for the 20 groups. A data set was created containing the 15 indicators for each case in the sample ($n = 5000$) for each group ("UD_MI_AM.dat"). The alignment method is easily implemented on Mplus using the INPUT INSTRUCTIONS below ("UD_MI_AM.inp"). Mplus requires the name of the variable containing the group ids as well as the total number of groups (classes). Then a unidimensional factor model is specified. Mplus produces some useful plots to visualise, in this case, the IRT parameters of the model. This is similar to Figures 5.1 to 5.4.

```
INPUT INSTRUCTIONS

  Data:
    File is UD_MI_AM.dat ;
  Variable:
    Names are id V1-V16 data;
    Missing are all (-9999) ;

  usevariables = V1-V15;

  categorical = V1-V15;

  classes = c(20);
  knownclass = c(data = 1 2 3 4 5 6 7 8 9 10
```

```
                        11 12 13 14 15 16 17 18 19 20);


 Analysis:  type = mixture;
           estimator = ml;
           alignment = free;
           ALGORITHM=INTEGRATION;
           Process=8;

 model:
           %overall%
           f by V1-V15;


   output:
           tech1 tech8 align;

   plot:
           type = plot2;
```

The results of the alignment method are displayed below. In rows the output indicates whether the parameter is variant or non-invariant (in brackets). The number corresponds to the number of the group. These findings suggest that all the thresholds are invariant and that all the loadings of items 5 to 15 (V5 to V20) are invariant. The loadings of the first four items are non-invariant. These results are consistent with the findings of the MGFA. However, in this case the analysis is performed for the 20 groups and only one model needed to be fitted to the data. The alignment method indicates that for these data is possible to minimize non-invariance if the loadings of items 1 to 4 (V1 to V4) are not fixed for groups 11 to 20. Therefore, the results suggest that partial scalar invariance holds for this simulated example.

The alignment method suggest that the 20 poverty indices are comparable once the non-invariance of the loadings of the first four items is accounted by for.

```
APPROXIMATE MEASUREMENT INVARIANCE (NONINVARIANCE) FOR GROUPS

 Intercepts/Thresholds
   V1$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V2$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V3$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V4$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V5$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V6$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V7$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V8$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V9$1        1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V10$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V11$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V12$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V13$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V14$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   V15$1       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

 Loadings for F
   V1          1 2 3 4 5 6 7 8 9 10 (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)
   V2          1 2 3 4 5 6 7 8 9 10 (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)
   V3          1 2 3 4 5 6 7 8 9 10 (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)
```

```
V4          1 2 3 4 5 6 7 8 9 10 (11) (12) (13) (14) (15) (16) (17) (18) (19) (20)
V5          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V6          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V7          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V8          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V9          1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V10         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V11         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V12         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V13         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V14         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
V15         1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

5.4 Real-data analysis of Measurement Invariance

Strong (or Scalar) measurement invariance is unlikely to hold when working with real data. The analysis of MI is fairly new in poverty research that there is no agreement about what the desirable level of MI should be.@Guio2017 found partial scalar MI using the EU-SILC data and Najera (2016) found partial scalar MI for the official Mexican multidimensional measure. Both exercises conclude that the ideal is scalar MI but that in practice partial scalar is a more sensible standard. Both non-invariant thresholds and loadings are likely to appear in real-data analyses but given that differences in difficulty -thresholds- could be more explicitly accounted by for in scale equating -next chapter- one sensible recommendation is to maximize metric invariance and then move onto partial scalar invariance.

The INPUT INSTRUCTIONS of the MI analysis of the FRS data ("FRS_mplusprep.dat") are shown below. The data set contains all the 25 deprivation indicators, sampling weights (gross4) and the household id (id). There is data available for ten periods (2004/2005 to 2013/2104). However, four variables were replaced by another four in 2011. Therefore, the MI analysis is performed using the 17 common variables across both periods -next chapter concerns with the issue of equating and scaling measures with different and non-invariant indicators). The classes in the INPUT INSTRUCTIONS are the 10 years (ordered were 1=2004/2005, ..., 10=2013-2014). For the purposes of the illustration cluster sampling is assumed but we recommend checking the Mplus manual for deeper understanding of working with complex samples.

```
INPUT INSTRUCTIONS

  Data:
    File is FRS_mplusprep.dat ;
  Variable:
    Names are
       gross4 FRSYear dep_ADDDEC dep_ADDEPLES dep_ADDHOL
       dep_ADDINS dep_ADDMEL dep_ADDMON dep_ADDSHOE dep_ADEPFUR
       dep_AF1 dep_AFDEP2 dep_HOUSHE1 dep_CDELPLY dep_CDEPBED
       dep_CDEPCEL dep_CDEPEQP dep_CDEPHOL dep_CDEPLES
       dep_CDEPSUM dep_CDEPTEA dep_CDEPTRP dep_CPLAY
       dep_CDEPACT dep_CDEPVEG dep_CDPCOAT dep_ADBTBL id;
    Missing are all (-9999) ;

  usevariables = dep_ADDDEC  dep_ADDHOL dep_ADDINS
        dep_ADDMON  dep_ADEPFUR dep_AF1 dep_AFDEP2 dep_HOUSHE1
       dep_CDELPLY dep_CDEPBED dep_CDEPCEL dep_CDEPEQP dep_CDEPHOL
       dep_CDEPLES dep_CDEPTEA dep_CDEPTRP dep_CPLAY;

  categorical = dep_ADDDEC  dep_ADDHOL dep_ADDINS
        dep_ADDMON  dep_ADEPFUR dep_AF1 dep_AFDEP2 dep_HOUSHE1
```

```
      dep_CDELPLY dep_CDEPBED dep_CDEPCEL dep_CDEPEQP dep_CDEPHOL
      dep_CDEPLES dep_CDEPTEA dep_CDEPTRP dep_CPLAY;

  classes = c(10);
  knownclass = c(FRSYear = 1 2 3 4 5 6 7 8 9 10);

  weight=gross4;

  cluster=id;

  Analysis:  type = mixture complex;
             estimator = ml;
             alignment = fixed;
             ALGORITHM=INTEGRATION;
             Process=8;

  model:

             f by dep_ADDDEC  dep_ADDHOL dep_ADDINS
       dep_ADDMON  dep_ADEPFUR dep_AF1 dep_AFDEP2 dep_HOUSHE1
      dep_CDELPLY dep_CDEPBED dep_CDEPCEL dep_CDEPEQP
      dep_CDEPHOL dep_CDEPLES
      dep_CDEPTEA dep_CDEPTRP dep_CPLAY;


    output:
             tech1 tech8 align;

    plot:
             type = plot2;
```

One of the key outputs of the alignment method is displayed below and it shows which parameters are non-invariant (within brackets) for each year. For example, the intercept of DEP_ADDD non-invariant for group 1 and the intercepts of DEP_ADDH and DEP_ADDI are invariant. The results suggest that partial scalar MI holds. This can be achieved by using a subset of items with both invariant intercepts and loadings: DEP_ADDI, DEP_ADEP, DEP_AF1, DEP_AFDE, DEP_CDEPEQP, DEP_CDEPHOL, DEP_CDEPTRP and DEP_CPLA. There are other items that are almost fully invariant as they only violate MI in one group. At this point is worth noting that the alignment method finds the parameters that maximize approximate MI and, therefore, MGFA could be used to find more items that fulfill MI once other parameters have been fixed, this the advantage but also the disadvantage of the MGFA in that it allows the researcher to find that partial scalar MI models -if exists- that suits better the purposes of her research.

```
APPROXIMATE MEASUREMENT INVARIANCE (NONINVARIANCE) FOR GROUPS

 Intercepts/Thresholds
   DEP_ADDD$1  (1) 2 3 4 5 6 7 8 9 10
   DEP_ADDH$1  1 2 3 4 5 6 7 8 9 10
   DEP_ADDI$1  1 2 3 4 5 6 7 8 9 10
   DEP_ADDM$1  (1) (2) (3) (4) 5 6 7 8 9 (10)
   DEP_ADEP$1  1 2 3 4 5 6 7 8 9 10
   DEP_AF1$1   1 2 3 4 5 6 7 8 9 10
   DEP_AFDE$1  1 2 3 4 5 6 7 8 9 10
   DEP_HOUS$1  (1) (2) 3 4 5 6 7 8 9 10
   DEP_CDEL$1  (1) (2) (3) (4) (5) 6 7 8 9 10
   DEP_CDEP$1  1 2 3 4 5 6 7 8 9 10
```

```
   DEP_CDEP$1  1 2 3 4 5 6 7 8 9 10
   DEP_CDEP$1  1 2 3 4 5 6 7 8 (9) 10
   DEP_CDEP$1  (1) 2 3 4 5 6 7 8 9 10
   DEP_CDEP$1  1 2 3 4 5 6 7 8 9 (10)
   DEP_CDEP$1  1 2 3 4 5 6 7 8 9 10
   DEP_CDEP$1  1 2 3 4 5 6 7 8 9 10
   DEP_CPLA$1  1 2 3 4 5 6 7 8 9 10

 Loadings for F
   DEP_ADDD    (1) 2 3 4 5 6 7 8 9 10
   DEP_ADDH    (1) (2) (3) (4) 5 6 7 8 9 10
   DEP_ADDI    1 2 3 4 5 6 7 8 9 10
   DEP_ADDM    1 2 3 4 5 (6) 7 8 9 (10)
   DEP_ADEP    1 2 3 4 5 6 7 8 9 10
   DEP_AF1     1 2 3 4 5 6 7 8 9 10
   DEP_AFDE    1 2 3 4 5 6 7 8 9 10
   DEP_HOUS    1 2 (3) (4) 5 6 7 8 9 10
   DEP_CDEL    1 2 3 4 5 6 7 8 9 10
   DEP_CDEP    1 2 3 4 5 6 7 8 9 10
   DEP_CDEP    1 2 3 4 5 6 7 8 9 10
   DEP_CDEP    1 2 3 4 5 6 7 8 (9) 10
   DEP_CDEP    1 2 3 4 5 6 7 8 9 10
   DEP_CDEP    (1) 2 3 4 5 6 7 8 9 10
   DEP_CDEP    1 2 3 4 5 6 7 (8) (9) 10
   DEP_CDEP    1 2 3 4 5 6 7 8 9 10
   DEP_CPLA    1 2 3 4 5 6 7 8 9 10
```

One of the key purposes of the alignment method is to compare groups on the factor mean once approximate MI is met. The results suggest that the severity of child deprivation has remained pretty much the same. Years, 5 (2008/09), 6 (2009/10) and 7 (2010/11), however, have significantly smaller means compared with 2004/2005. This suggest that in the observed period and with these 17 indicators, child deprivation was at its lowest level in 2008/09 (pre-economic crisis). It seems that after the year 2010/11 severity increased and reached similar levels of the early 2000s.

```
FACTOR MEAN COMPARISON AT THE 5% SIGNIFICANCE LEVEL IN DESCENDING ORDER

 Results for Factor F

           Latent    Group      Factor
 Ranking    Class    Value       Mean     Groups With Significantly
                                          Smaller Factor Mean
     1         1        1       0.000    6 7 5
     2         4        4      -0.005    5
     3         3        3      -0.011    5
     4         9        9      -0.014    5
     5        10       10      -0.029
     6         8        8      -0.034
     7         2        2      -0.036
     8         6        6      -0.042
     9         7        7      -0.048
    10         5        5      -0.065
```

```
library(haven)
library(plyr)
```

```
D<-read_dta("FRS_mplusprep.dta")

if(D$FRSYear<=7) {
  D$ds_uw<-rowSums(D[,c(3:23)],na.rm=TRUE)
} else {
  rowSums(D[,c(3,5,6,8,10,11:19,21:27)],na.rm=TRUE)
}

ds_frs<-ddply(D,.(FRSYear), summarise, mean_ds=mean(ds_uw))
ds_frs$year<-c("2004/2005","2005/2006","2006/2007","2007/2008","2008/2009","2009/2010","2010/2011","201
                   "2012/2013","2013/2014")
frs_raw<-data.frame(ds_raw=c(0.000 , -0.005,  -0.011, -0.014, -0.029, -0.034, -0.036, -0.042, -0.048, -0
ds_frs<-merge(ds_frs,frs_raw,by="FRSYear")
```

Figure 5.5 plots the raw child deprivation score (0-17 count) and the adjusted factor score. When a scale is highly reliable and fully invariant, the raw mean and the factor score must be highly (negatively) correlated. The plot does not shows a clear correlation. The MI analysis suggested that several parameters need to be taken into account for the scale to be comparable over time. Whereas the raw deprivation score suggest that severity of deprivation has decreased over time (particularly after 2011), the MI analysis suggest that this is not the case as the scale lost comparability. Therefore, researchers relying on raw scores or any other transformation that does not takes into account non-invariance are likely to arrive to incorrect conclusions about the trends in child deprivation.

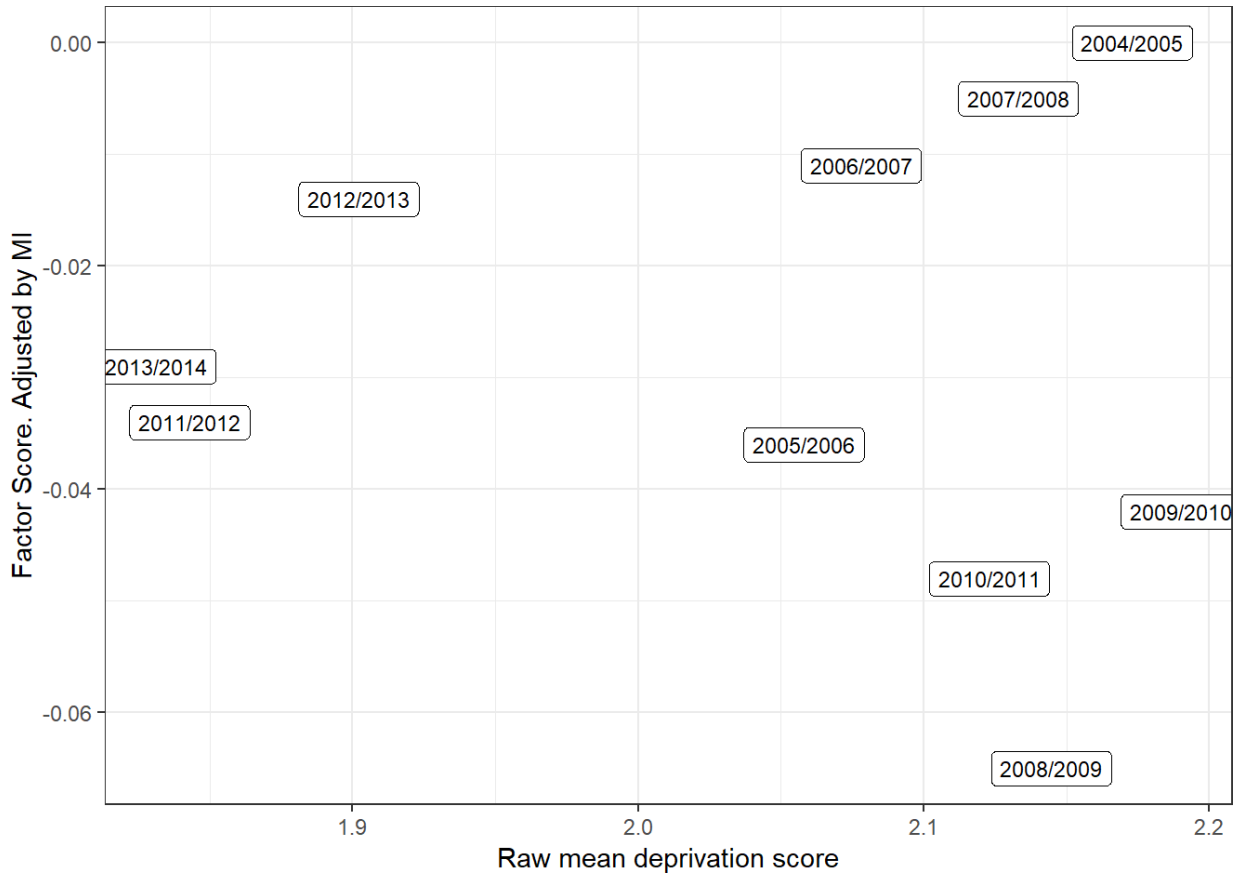

Figure 5.5: Scatter plot with a comparison of the raw mean child deprivation score (simple count) and the factor adjusted score. FRS data 2004-2013. The plot shows that there a great deal of discrepancy between the adjusted and the unadjusted severity scores. Conclusions about the depth of poverty are affected by the comparability of the data over time.

Prevalence weighting has been a means to consider the severity of item-level deprivation in the estimation of overall deprivation and poverty (**???**). This procedure consists in assigning more weight to those items that most people have in the society in question. As discussed in the chapter on Reliability, an index is self-weighting for high reliability values. Therefore, very little is gained when using differential weighting. However, when comparing groups or years weighting could help to improve the comparability of a measure as this procedure is directly associated with the difficulty/severity parameter. Adjusting by differences in severity is one way to improve comparisons across groups. The next chapter, therefore, focuses on how equating, linking and scaling can be used to make scales comparable.

Chapter 6

Scale equating and linking

6.1 Intuition to scale equation

Previous chapter introduces the sources affecting the comparability of different poverty indices and presents the concept and empirical implementation of measurement invariance. In poverty research, households are ranked according to some indicators of (low) living standards using survey or census data. The indicators in question have a structure (unidimensional or multidimensional, See chapter X) for the population in question. One problem is that such a measurement model might not be adequate for a different population or for a different year. That is, the measurement model is not equivalent (see previous chapter). Therefore, we would like to assess whether a measure is equivalent across populations/periods. Once this assessment is conducted poverty can be compared on the factor using the aligment method. This, however, might not be fully satisfactory for policy makers as the values of a standarized latent variable make little practical sense. Furthermore, it is unclear how to use these values to set a poverty line.

Another critical problem is that MI is adequate when scales have the same items and it does not solve the problem of working with scales that have different items or that have been upgraded in accordance with the living standards. Ideally, once measurement invariance is assessed researches would like to put everything into a meaninful metric and being able to compare measures that might have suffered from changes in its contents. Moreover, one question in poverty research is about how the severity of poverty is affected by changes in living standards and how this can be tractable using the available data.

6.2 Theory of scale equating

6.3 Example with simulated data in R

6.4 Real-data example

Chapter 7

Identifying the poor group

Chapter 8

Final thoughts

8.1 The future of data production in multidimensional poverty measurement

8.2 Advanced topics in multidimensional poverty measurement

Chapter 9

References

AERA, APA and NCME. (2014). Standards for educational and psychological testing. (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), Ed.). Amer Educational Research Assn.

Alkire, S. (2007). Choosing dimensions: The capability approach and multidimensional poverty. In *The many dimensions of poverty* (pp. 89–119). Springer.

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

Alkire, S., & Roche, J. (2011). *Beyond headcount: Measures that reflect the breath and components of child poverty* (No. 2). Oxford Poverty; Human Development Initiative (OPHI).

Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E., & Seth, S. (2015). *Multidimensional poverty measurement and analysis*. Oxford University Press, USA.

Alkire, S., & Santos, M. (2010). *Acute multidimensional poverty: A new index for developing countries*. OPHI Working Paper No. 38.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Altimir, O. (1979). *La dimensión de la pobreza en América Latina*. CEPAL.

Atkinson, A. B., Guio, A.-C., & Marlier, E. (2017). *Monitoring social inclusion in europe*. Publications Office of the European Union.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Bartholomew, D. J. (1987). *Latent variables models and factor analysis.* (D. J. Bartholomew, Ed.). New York: Oxford University Press.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829. doi:http://dx.doi.org/10.1016/j.paid.2006.09.024

Betti, G., Gagliardi, F., Lemmi, A., & Verma, V. (2015). Comparative measures of multidimensional deprivation in the european union. *Empirical Economics*, *49*(3), 1071–1100.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Boltvinik, J. (1992). El método de medición integrada de la pobreza. Una propuesta para su desarrollo. *Comercio Exterior*, *42*(4), 354–365.

Boltvinik, J. (1998). *Poverty measurement methods: An overview*. Series on Poverty Reduction: An overview.

Boltvinik, J. (2014). América Latina, de la vanguardia al rezago en medición multidimensional de la pobreza. La experiencia contrastante de México . Una guía para la región. In J. et a. Boltviniki (Ed.), *La multidimensionalidad como un desafío para los métdos y técnicas de la medición de la pobreza.* CLACSO-CROP.

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

Bradshaw, J. (1993). *Budget standards for the united kingdom.* (J. Bradshaw, Ed.). Aldershot: Avebury.

Bradshaw, J., Holmes, H., & Hallerod, B. (1995). Adapting the consensual definition of poverty. *Breadline Britain in the 1990s, Department of Social Policy and Planning, University of Bristol, Bristol.*

Bradshaw, J., Middleton, S., Davis, A., Oldfeld, N., Smith, N., Cusworth, L., & Williams, J. (2008). *A minimum income standard for britain.* Creative Commons. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/3465/1/2226-income-poverty-standards.png

Brennan, R. L. (2006). *Educational measurement. ACE/praeger series on higher education.* ERIC.

Brown, T. (2006). *Confirmatory factor analysis for applied research.* (T. Brown, Ed.). The Guilford Press.

Browne, M. W., Cudeck, R., & others. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *154*, 136–136.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456.

Clark, D. A., & Qizilbash, M. (2005). Core poverty, basic capabilities and vagueness: An application to the south african context.

CONEVAL. (2011a). *Medición de la pobreza en los municipios de México.* CONEVAL.

CONEVAL. (2011b). Metodología para la medición multidimensional de la pobreza en México. *Realidad, Datos Y Espacio. Revista Internacional de Estadística Y Geografía*, *2*(1), 36–63.

Cortés, F. (2014). La medición multiimensional de la pobreza en México. In J. et a. Boltvinik (Ed.), *La multmultidimensional como un desafío para los métdos y técnicas de la medición de la pobreza.* CLACSO-CROP.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

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

Decancq, K., & Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, *32*(1), 7–34. doi:10.1080/07474938.2012.690641

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, *52*(3), pp. 761–766. Retrieved from http://www.jstor.org/stable/1913475

Foster, J., Greer, J., & Thorbecke, E. (2010). The foster–greer–thorbecke (fgt) poverty measures: 25 years later. *The Journal of Economic Inequality*, *8*(4), 491–524. doi:10.1007/s10888-010-9136-1

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research*, *25*(3), 201–239.

Gibbons, R. D., Immekus, J. C., Bock, R. D., & Gibbons, R. D. (2007). The added value of multidimensional irt models. *Multidimensional and Hierarchical Modeling Monograph*, *1*.

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 Policy Press.

Gordon, D. (2010). Metodología de medición multidimensional de la pobreza a partir del concepto de privación relativa. In M. Mora (Ed.), *La medicion de la pobreza multidimensional en México* (pp. 401–498). El Colegio de México. CONEVAL.

Gordon, D. (2018). Measuring poverty in the uk. In E. Dermott & G. Main (Eds.), *Poverty and social exclusion in the uk*.

Gordon, D., & Nandy, S. (2012). Measuring child poverty and deprivation: Measurement, concepts, policy and action. In *Global child poverty and well-being* (pp. 57–102). The Policy Press. University of Bristol.

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

Guio, A.-C. (2009). *What can be learned from deprivation indicators in europe*. eurostat. Methodologies; working papers.

Guio, A.-C., Gordon, D., Marlier, E., Najera, H., & Pomati, M. (2017). Towards an eu measure of child deprivation. *Child Indicators Research*. doi:10.1007/s12187-017-9491-6

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. doi:10.1177/0958928716642947

Guio, A., Fusco, A., & Marlier, E. (2009). *A european union approach to material deprivation using eu-silc and eurobarometer data*. International Networks for Studies in Technology, Environment, Alternatives; Development.

Guio, A., Gordon, D., & Marlier, E. (2012). *MEASURING material deprivation in the eu: Indicators for the whole population and child-specific indicators*. EUROSTAT.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. doi:10.1007/BF02288892

Halleröd, B. (1995). The truly poor: Direct and indirect consensual measurement of poverty in sweden. *Journal of European Social Policy*, *5*(2), 111–129. doi:10.1177/095892879500500203

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, 1–18. doi:10.1080/10705511.2017.1402334

Harris, D. (1989). Comparison of 1, 2, and 3-parameter irt models. *Educational Measurement: Issues and Practice*, *8*(1), 35–41. doi:10.1111/j.1745-3992.1989.tb00313.x

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Joreskog, K. G., Sorbom, D., & Magidson, J. (1979). Advances in factor analysis and structural equation models.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *ETS Research Bulletin Series*, *1970*(2), i–45.

Kakwani, N., & Silber, J. (2008). *Many dimensions of poverty*. Springer.

Katzman, R. (2000). *Notas sobre la medición de la vulnerabilidad social*. CEPAL.

Klasen, S. (2000). Measuring poverty and deprivation in south africa. *Review of Income and Wealth*, *46*(1), 33–58.

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. doi:10.1002/cem.2427

Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method.

Lazardfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. In. Boston: Houghton Mifflin.

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, *7*(1), 84.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*(6), 543–566. doi:http://dx.doi.org/10.1016/S0160-2896(03)00051-5

Lumley, T. (2011). *Complex surveys: A guide to analysis using r* (Vol. 565). John Wiley & Sons.

Lumley, T. (2016). Survey: Analysis of complex survey samples.

Mack, J., & Lansley, S. (1985). *Poor britain*. (J. Mack & S. Lansley, Eds.). London, George Allen & Unwin.

Martinetti, E. C. (2006). Complexity and vagueness in the capability approach: Strengths or weaknesses. *University of Pavia, Italy,(Mimeo)*.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370.

Max-Neef, M., Elizalde, A., & Hopenhayn, M. (1992). Development and human needs. *Real-Life Economics: Understanding Wealth Creation*, 197–213.

McCullagh, P. (2002). What is a statistical model? *Annals of Statistics*, 1225–1267.

McDonald, R. P. (1999). *Test theory: A unified treatment*. (R. P. McDonald, Ed.). Mahwah, N.J. L. Erlbaum Associates.

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

McKay, S. (2004). Poverty or preference: What do consensual deprivation indicators really mean? *Fiscal Studies*, *25*(2), 201–223.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), pp. S69–S77. Retrieved from http://www.jstor.org/stable/41219507

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

Michell, J. (2015). Measurement theory: History and philosophy. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences (second edition)* (Second Edition., pp. 868–872). Oxford: Elsevier. doi:https://doi.org/10.1016/B978-0-08-097086-8.43062-X

Mora, M. (2010). Medición multidimensional de la pobreza en México. In M. Mora (Ed.), *Medición multidimensional de la pobreza en México* (pp. 1–25). El Colegio De México, CONEVAL.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi:10.1007/BF02294210

Muthén, B. (2007). Laten variable hybrids. Overview of old and new models. In G. Hancock & K. Samuelsen (Eds.), *Advances in latent variable mixture models*. Information Age Publishing.

Muthén, B. (2013). *IRT in mplus*. Mplus. Retrieved from http://www.statmodel.com/download/MplusIRT2.png

Muthén, L., & Muthén, B. (2012). *Mplus user's guide. Seventh edition*. (L. Muthén & B. Muthén, Eds.). Mplus.

Najera, H. E. (2016). Does measurement invariance hold for the official mexican multidimensional poverty measure? A state-level analysis 2012. *Quality & Quantity*, 1–25. doi:10.1007/s11135-016-0327-0

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. doi:10.1007/s11205-014-0819-z

Narayan, D. (2001). Voices of the poor. *Faith in Development: Partnership Between the World Bank and the Churches in Africa, Washington, DC: World Bank and Oxford: Regnum Books*, 39–50.

Nájera, H. (n.d.). Scale reliability and the monotonicity axiom in multidimensional poverty measurement. *Social Indicators Research*.

Nájera, H. E. (2018). Reliability, population classification and weighting in multidimensional poverty measurement: A monte carlo study. *Social Indicators Research*. doi:10.1007/s11205-018-1950-z

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. doi:https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J., & Bernstein, I. (1994). In *Psychometric theory* (3rd ed.). McGraw Hill.

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

Pantazis, C., Gordon, D., & Levitas, R. (2006). *Poverty and social exclusion in britain: The millennium survey*. (C. Pantazis, D. Gordon, & R. Levitas, Eds.). Policy Press. Retrieved from https://books.google.com/books?id=o-H0J4BMWS8C

Pogge, T. (2005). World poverty and human rights. *Ethics & International Affairs*, *19*(1), 1–7.

Ravallion, M. (2010). A reply to reddy and pogge. In S. Anand, S. P., & J. Stiglitz (Eds.), *Debates on the measurement of global poverty* (p. p. 43). Oxford Scholarship Online.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Reddy, S., & Pogge, T. (2010). How not to count the poor. In S. Anand, S. P., & J. Stiglitz (Eds.), *Debates on the measurement of global poverty* (p. p. 43). Oxford Scholarship Online.

Reise, S. P. (2014). Item response theory. In *The encyclopedia of clinical psychology* (pp. 1–10). American Cancer Society. doi:10.1002/9781118625392.wbecp357

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.* R software.

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, *74*(1), 145–154. doi:10.1007/s11336-008-9102-z

Rigdon, E. E. (1996). CFI versus rmsea: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(4), 369–379. doi:10.1080/10705519609540052

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rowntree, S. (1901). *Poverty: A study of town life*. (S. Rowntree, Ed.). Macmillan; Co.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, *54*(2), 189–203. doi:https://doi.org/10.1016/j.im.2016.06.005

Santos, M. E., & Villatoro, P. (2016). A multidimensional poverty index for latin america. *Review of Income and Wealth*, n/a–n/a. doi:10.1111/roiw.12275

Schoot, R. van de, Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi:10.1080/17405629.2012.686740

Sen, A. (1976). Poverty: An ordinal approach to measurement. *Econometrica*, *44*(2), pp. 219–231. Retrieved from http://www.jstor.org/stable/1912718

Sen, A. (1983). Poor relatively speaking. *Oxford Economic Papers*, *35*, 153–169.

Sen, A. (1985). A sociological approach to the measurement of poverty: A reply to peter townsend. *Oxford Economic Papers*, *37*, 669–676.

Sen, A. (2005). Human rights and capabilities. *Journal of Human Development*, *6*(2), 151–166. doi:10.1080/14649880500120491

Skrondal, A., & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi:10.1111/j.1467-9469.2007.00573.x

Spearman, C. (1904). The proof and measurement of association between two things. *Américan Journal of Psychology*, *15*(1), 72–101.

Spicker, P., Alvarez, S., & Gordon, D. (2006). *Poverty and international glossary.* (P. Spicker, Alvarez S., & D. Gordon, Eds.). International Studies in Poverty Research. International Social Science Council. Zen Books.

Steiger, J. H. (1980). Statistically based tests for the number of common factors. In *The annual meeting of the psychometric society. Iowa city, ia. 1980.*

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use.* Oxford University Press, USA.

Thorbecke, E. (2007). Multidimensional poverty: Conceptual and measurement issues. In *The many dimensions of poverty* (pp. 3–19). Springer.

Thorndike, R., & Hagen, E. (1969). Measurement and evaluation in education and psychology. New York, NY: John Wiley; Sons.

Thurstone, L. (1947). Multiple factor analysis. In. University of Chicago Press.

Townsend, P. (1979). *Poverty in the united kingdom: A survey of household resources and standards of living.* (P. Townsend, Ed.). University of California.

Townsend, P. (1985). A sociological approach to the measurement of poverty–a rejoinder to professor amartya sen. *Oxford Economic Papers*, *37*(4), pp. 659–668. Retrieved from http://www.jstor.org/stable/2663048

Townsend, P. (1987). Deprivation. *Journal of Social Policy*, *16*(02), 125–146. doi:10.1017/S0047279400020341

Townsend, P. (1993). The politics of poverty and health. *BMJ: British Medical Journal*, *306*(6873), p. 337. Retrieved from http://www.jstor.org/stable/29718417

Townsend, P., & Gordon, D. (1993). How much is enough? In P. Townsend (Ed.), *The international analysis of poverty* (pp. 1–1). Harvester Wheatsheaf.

Townsend, P., & Gordon, D. (2000). *Breadline europe: The measurment of poverty.* (P. Townsend & D. Gordon, Eds.). Bristol Policy Press.

Tsui, K.-y. (2002). Multidimensional poverty indices. *Social Choice and Welfare*, *19*(1), 69–93.

UNDP. (2014). *Multidimensional Poverty Index (MPI).* UNDP.

Van den Bosch, K. (2001). *Identifying the poor: Using subjective and consensual measures.*

Whelan, C., Nolan, B., & Maitre, B. (2006). Measuring consistent poverty in ireland with eu silc data.

World-Bank. (2017). Monitoring global poverty: Report of the commission on global poverty (the atkinson commission). *World Bank, Washington.*

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. doi:10.1007/s11336-003-0974-7