TITLE

Normative validity: the case of poverty measures[1]

This manuscript is a work in progress and has been accepted for presentation at the Philosophy of Science Conference, *PSA 2023 Around the World*. **Please do NOT cite or quote it.**

Author: Samuel Maia

Affiliation: Department of Philosophy, Federal University of Minas Gerais (UFMG)

ORCID: 0000-0003-0430-0767

E-mail: samuelmaia@ufmg.br

Site: www.samuelmaiabr.com

ABSTRACT

Drawing from Anna Alexandrova's (2017) work on the measurement of well-being, I expand on what she termed "normative validity," using poverty measures as a case study. Normative validity refers to a measure's ability to accurately capture the moral and political aspects of its target concept. Initially, I outline its characteristics, comparing it with data-driven validity and discussing how it fits in measurement theory. I then show how researchers in poverty measurement employ various tools to ensure their measures capture the normative elements of this concept. I conclude by exploring how normative validity sheds light on the tension between normative and data-driven poverty measures.

[1] I presented portions of this paper at the Workshop on Thick Concepts in Philosophy of Science at Bielefeld University, the Philosophy of Social Science Roundtable 2021 at the University of Nebraska, and the Cambridge Early Career Workshop in the Philosophy of Measurement at the University of Cambridge. I extend my gratitude to the organizers of these venues and appreciate the invaluable feedback from Daniel Katz, Cristian Larroulet Philippi, Alex Mussgnug and Eran Tal. Telma Birchal and Mateus Leite read earlier versions of this draft. I am especially indebted to Anna Alexandrova for her insightful and challenging comments on my arguments here. I stand by the content, acknowledging any errors or omissions that may persist.

> *"'Normative Validity' … would address whether or not the measure arrived at by the standard methods … does not have obvious problems from a normative point of view" (Alexandrova, 2017: 151).*
>
> *"The measurement of poverty is not a purely technical subject … [it]is not like a guide to plumbing, because the right answers depend on views that are politically influenced and, at heart, matters of moral judgment" (Atkinson, 2019: 212).*

INTRODUCTION

This paper develops an account of normative validity and illustrates it through poverty measures. Many ideas addressed below are legatees of Anna Alexandrova's reflections on what she called "value-aptness" in measuring well-being. To my knowledge, she coined the term "normative validity." Still, my goal is to treat normative validity more generally than she did and, for that matter, to say something useful for researchers who work with concepts other than well-being – like poverty but not only.

I start with a definition of normative validity:

- ***Normative validity:*** the property of a measure that adequately captures its targeted concept's normative (moral and political) aspects.

Part 1 starts with an outline of the general features of this definition *(1.I)*. Then, I discuss the accounts of validity from a version of the representational theory of measurement and psychometrics and explain why normative validity can be better understood within their framework (*1.II*). Normative validity is conceptually-driven, while other commonly recognized types of validity are data-driven (*1.III*). I detail how normative validity can be achieved (*1.IV*) and discuss its objections (*1.V*). Part 2 shows how normative validity is pursued in poverty measurement, even without being explicitly called so (*2.I*). Researchers employ informal and formal tools when identifying individuals in poverty (*2.II.a*) or creating a poverty index (*2.II.b*) to argue that a poverty measure captures the normative features of its concept. Lastly, I explore how normative validity can help us better understand the conflict between normative and data-driven poverty measures (*2.III*).

PART 1: AN ACCOUNT OF NORMATIVE VALIDITY

1.I. Normative *and* validity

Normative validity (NV) is a *type* of validity. Therefore, to fully understand NV, we need to understand what validity in measurement is. Before this, I will clarify what I mean by a concept's normative aspects with a brief note on the type of concepts NV refers to.

1.I.a. Thick concepts in philosophy of science and in poverty measurement

Thick concepts are a hot topic in philosophy of science. They are part of the literature that debates the value-free ideal and its many ramifications, such as: whether the sciences and its products are carried out without non-epistemic values; if they are not, whether this would be possible; and if impossible, whether pursuing it would still be desirable.

Challenging this ideal, today's philosophers of science have been identifying different *loci* of values' influence in science (Elliot, 2017). There is almost a consensus that *thick concepts* constitute one of these places, as they jointly describe and reflect normative judgments. By playing central roles in theories, models, and *measures*, all these products of science can also be influenced by values (Alexandrova & Fabian, 2022: 7n1). Furthermore, some philosophers have argued that getting rid of values in those concepts is impossible or that, even if it is possible, this would not be justifiable (Reiss, 2022: 229). This paper will not offer other arguments nor discuss the metaethical nature of thick concepts.

Here, I will adopt a definition of thick concepts based on an adaptation of Elizabeth Anderson's (2002: 504-505):

- ***Thick concept:*** a concept whose application is guided (*a*) by empirical facts as well as (*b*) evaluative judgments, and (*c*) it licenses prescriptive judgments.

Concepts like well-being, inequality and of course poverty are thick. Let us see how this is the case of the concept of poverty. For instance, poverty is a kind of *deprivation*. But what kind of deprivation? What *thing* are people in poverty deprived of? Answering this translates into defining the concept's *deprivation space*, *i.e.*, choosing the type of deprivation a person in poverty faces. Some of the most famous contenders are economic resources, well-being, basic needs, a society's

standard of living and basic capabilities. For many poverty researchers (Spicker, 2006; Lister, 2021), choosing one may be an evaluative (moral or political) judgment as we may think that beyond an empirical judgment (*e.g.*, that a person's income is below a given threshold) poverty is a type of unacceptable deprivation or a matter of distributive justice. Also, one may claim that saying that a set of people is in poverty entails that a prescriptive judgment, *e.g.*, we need to do something to help them overcome poverty (Piachaud, 1981; Mack & Lansley, 1985). So, the application of poverty would be guided by evaluative judgments and would also license prescriptive judgments.[2]

However, from a concept being thick, it does not follow its measure will be. I will address this further on. Now, it is time to talk about validity.

1.I.b. Validity

Now, a definition of validity:[3]

- ***Validity:*** the property of a measure that adequately represents its targeted concept.

Let me mention important features of validity. First, we can distinguish validity from *validation*. The former refers to a *property* of measures that are in the right relation with what it intends to measure. The latter is a *process* through which someone provides *evidence* that a given measure is valid.[4] I will use both here. Still, I will focus on validation procedures, which put some

[2] There is a distinct, yet related, understanding of thick concepts that, although less common in the philosophy of science literature, is extensively discussed in the metaethical literature. In this context, a thick concept is an evaluative concept that stands in contrast with thin concepts. Standard thin evaluative concepts include terms like good/bad or right/wrong. These are viewed as highly abstract and possessing less (or none at all) descriptive content compared to thick concepts. Such abstract concepts can be applicable across different historical and cultural contexts, thus aligning with universalist views of ethics. Conversely, thick concepts are often more localized and specific to a culture, history, or society. Therefore, to truly understand thick concepts, one must grasp the unique characteristics of a particular society. As expected, thick concepts are associated with relativist or contextualist views of ethics. Proponents, such as Bernard Williams (1985), use them to challenge ethical endeavors that seek overarching, universal applicability. Although I believe this understanding of thickness might have relevance for the philosophy of science (and poverty measurement!), I will not delve into that topic in this paper.

[3] Validity concerns a relationship between a *measure* and a *concept*. Nonetheless one might claim that a measure aims at some *phenomena*. Definitions in references in the social sciences disagree: either one claims that a measure aims at a concept (Babbie, 2014: 148-52) or, like Zeller and Carmines (1980: 77-91) and Bollen (1989: 184-94), relate validity, first, to the phenomena of interest, but soon after assume that these phenomena are captured through concepts, and that these are the aims of validity.

[4] Authors like Messick (1989) have claimed that validity provides evidence for inferences and actions we might take based on a measurement. Adapting it, this fits with feature *c* of our definition of thick concepts. Nevertheless, his account of validity is contentious (Mari *et al.*, 2021: 90).

flesh into NV implications. Second, researchers often present enhanced versions of old measures, arguing that their new version is more valid than its predecessor. This suggests that validity is a matter of *degree*, which I submit. Third, I have mentioned that NV is a type of validity. Different types of validity look at different *dimensions* of a measure's validity. Different validation processes – which correspond to different *types* of validity – provide different types of evidence for a measure's validity. A measure may have a higher degree of validity in one dimension and less in another. Fourth, I agree with Larroulet Philippi (2020), who claims that a measure's validity should be assessed according to the specific research *purpose* the measure is put in place to satisfy. In scientific and political endeavors that employ scientific products like measures, there is not one legitimate purpose but many.

Therefore, when assessing a measure, we should look for a higher degree of the type of validity that is more relevant to its purpose. Some purposes will privilege the features of the *concept* to be measured. For thick concepts, one may look for a higher degree of NV, privileging a measure's capacity to represent its normative features. However, other purposes will make us privilege a measure's *explanatory/predictive power*. In the last section (sec. *2.III*), we will see how different purposes motivate different poverty measures.

1.II. Psychometrics and Representational Theory of Measurement

It is useful to place validity within two broader theoretical frameworks of measurement, psychometrics and the Representational Theory of Measurement (RTM). Psychometrics has arguably developed the most full-fledged account of validity (*cf.* Borsboom, 2005). This certainly explains why an important strand in the philosophical and methodological literature on validity in the social sciences takes psychometrics as their reference.[5] Yet, I think a specific version of RTM is also fruitful for validity in general and NV in particular. Nonetheless, I will start with psychometrics' richer validity typology.

In this literature, we find dozens of types of validity. Here, I will work with Robert Adcock and David Collier's (A&C) influential account of validity (2001), which, though devised for political science, is based on psychometrics. Their account has two desirable features: it pays due

[5] Following Adcock & Collier (2001: 530), in this paper, I am taking psychometrics as a tradition that includes psychological literature and educational testing.

respect to works about validity devised for other social sciences (Carmines & Zeller: 1979; Bollen: 1989; Babbie: 2014);[6] it is parsimonious, clustering different types of validity into the following three:

- ***Content validity:*** the property of a measure that adequately represents the content of its targeted concept.
- ***Convergent/discriminant validity:*** the property of different measures whose results exhibit empirical association with the same or different concepts. When the association of different measures of the same concept is high, we have *convergent validity*, which gives evidence that these measures are measuring the same concept; while when the association of different measures of the same concept is low, we have *discriminant validity*, which gives evidence these measures are measuring different concepts.[7]
- ***Nomological/construct validity:*** the property of a concept's measure that is compatible with reasonably well-established causal hypotheses about a concept.

This vocabulary of psychometrics is undoubtedly valuable for us as it makes explicit the different types of evidence one can mobilize when claiming a measurement's validity. We should note that the distinction between convergent/discriminant and nomological/construct validity maps the distinction between association and causation. The former relates to the matching between the empirical findings of a measure and the "descriptive" expectations among concepts and their components (2001: 540-41). The latter confirms or not whether a causal hypothesis is well established. A&C describes the reasoning behind nomological validity: "Assume the Hypothesis, Evaluate the Measure" (542).[8]

Now, let us discuss RTM. According to it, measurement examines and designs relationships (or mappings) between an empirical and a numerical structure. This examination and design are formal – *i.e.*, proofs of representation theorems that establish the conditions (axioms) for such

[6] A&C's account covers MacPhail's (1998) validity framework for economics and Goldthorpe's (2015: 64) commentaries on validity in sociology.
[7] When a measure's results converge with those of one or more measures taken as a standard of reference, we have *criterion validity*, a special case of convergent validity.
[8] As Alkire *et al.* (2015: 86-90) note, typical convergent/discriminant validity's methods are cluster analysis, Principal and Multiple Correspondence Analysis. Typical nomological/construct validity's (or *model-based*) methods are Latent Class Analysis, Factor Analysis, and Structural Equation Models.

mapping exist (Krantz *et al.*, 1971: 9).[9] Following Vessonen (2017) and Kahneman *et al.* (1997), I think RTM is a theory of measurement compatible with psychometrics. Moreover, I think RTM focuses on some aspects of validity that psychometrics neglects.

This is clear if we consider Nancy Cartwright and Norman Bradburn's (2011: 53-6) nonaxiomatic version of RTM. According to it, a good measure has three elements and arguments that these three are consistent and mutually supporting:

- A *characterization* of the concept to be measured;
- A formal *representation* of it;
- *Procedures* for assigning values to the items measured.

Let me put forward an interpretation of this theory, RTM*. While psychometrics provides a more detailed way to validate a measure, RTM* provides a framework for thinking about validity at a higher level of abstraction. Arguing for consistency and mutual support among characterization, representations and procedures is validating a measure.

This is how RTM* relates to what A&C (remember, a psychometrics-based account of validity) call measurement levels and tasks (2001: 530-531). In the social sciences, the concepts that animate our measures often are what Cartwright and Bradburn call *Ballung* concepts: concepts with multiple meanings and fuzzy boundaries, and that are used as typical currency in everyday speech. This is closer to the *background concept* (level 1) in A&C's terms. To figure in scientific theories or measures, we must precisify this concept – *i.e.*, rid it of its many meanings and features, narrowing down its extensionality and sometimes revising its folk meaning and extension. This corresponds to characterization, and its product is called the *systematized concept* (level 2) by A&C, or what some psychometricians call a *construct*. Now, we may interpret content validity as providing arguments for a concept's specific characterization and showing that its representation, which A&C call *indicators* (level 3), is adequate for it. The outputs of procedures, called by A&C as indicator *scores* (level 4), provide the empirical ingredient for the statistical and theoretical analysis carried out in convergent/discriminant and nomological/construct validation. I take this remarkable similarity between the structure of Cartwright and Bradburn's theory and A&C's measurement levels and tasks as evidence that RTM* and psychometrics are compatible, and both express something about validity.

[9] Or for such a mapping to be *valid*.

According to classical RTM, we must provide formal arguments – representation theorems and their proofs – in which the precisified concept and their representation are in the right relationships. However, RTM* emphasizes that informal arguments may be for this validation process. We shall see below (sec. *2.II*) how normative validation in poverty measurement take place with formal and informal arguments.

1.III. Two foci of validity

In his classic on social science measurement, Kenneth Bollen (1989: 185, 206) classifies our three types of validity into two groups with different focuses. Convergent/discriminant and nomological/construct validity focus on the measure *scores* or outputs, which constitute evidence for validation after being interpreted through statistical tools. In contrast, content validity focuses on the *concept* that a measure intends to capture. Here, evidence for validation is given through *conceptual or theoretical arguments* that a measure represents the concept's meaning and dimensions.

They are, respectively, *data-driven*, privileging a measure's explanatory/predictive power, and *conceptually-driven*, privileging a measure's fit with conceptual or theoretical perspectives (Bollen, 1989: 185).[10] In practice, the same measure may enjoy different validity statuses, depending on whether we are privileging data-driven or conceptually-driven validity. Also, when designing a measure, a researcher may face a trade-off on what will be her focus – explanatory power or conceptual adequacy. It is true that theoretical considerations are important for both groups. Yet, the focus of data-driven approaches lies in the empirical domain, specifically on how well the data aligns with theory. This is underscored by their advocates who champion statistical methods as a "falsifiable methodology." Conversely, in conceptually-driven approaches, it's primarily the theoretical (and normative) factors that dictate which data elements will shape our measurements.

[10] Here, I loosely use explanation to include causal relations but also to capture what Townsend – an illustrious champion of data-driven poverty measures – conveys when arguing that good characterizations "should involve the ordering of a mass of factual data in a rational, orderly and informative fashion" (1979: 38). Also, conceptually-driven validity can pay attention to a measure outputs. However, while in data-driven validities the outputs are evidence for conceptual revision, in conceptually-driven validities the outputs that should accord with conceptual expectations – or, in the case of axiomatic poverty measures, normative *dicta*.

As a subtype of content validity, NV focuses on the concept side of validation. It is *normative-driven*, as it emphasizes the adequacy of a measure to its concept's normative aspects. Its relevant evidence is not statistical but rather theoretical or conceptual arguments about normative matters.

We will see how NV may conflict with data-driven validities in poverty measurement.

1.IV. How to achieve normative validity?

How to achieve NV? Or, to put it in another way, what consists of NV validation? What does it require?

There are two types of normative validation. In the first, the *loci* of normative validation are the normative (moral and political) judgments that a measure designer faces when designing a measure based on a thick *Ballung* concept. A measure is a product of choices. A measure of thick concepts is the product, *inter alia*, or normative judgments (thick concept's element *b*). In the second, normative validation assesses whether the outputs of the normative judgments – the three outputs of RTM*'s steps: concept, representation, and procedures – are consistent and in the right relations.

Does it mean that NV will be a function of the truthness or rightness of these moral and political judgments? This would be undesirable, as measurement would depend excessively on contentious issues. I prefer to present some *desiderata* for these choices and claims that the normative judgment outputs are consistent and in the right relation. They are:[11]

(1) transparency about the *loci* of normative judgments;

(2) articulation and defense against alternatives of the *content* and *source* of someone's normative choices.

Note that *desideratum 2* concerns the content of a normative judgment and its source. By source, I mean the different types of information measurement that designers may appeal to when making these judgments. Adapting Alkire *et al.*'s (2015: 202-06) list, we have the following sources: (a) deliberative or participatory exercises; (b) representative surveys; (c) enduring

[11] This account is inspired by Alexadrova's rules for procedural objectivity in measuring well-being (2017: chapter 4). In fact, I interpret her account of assuring objectivity as providing part of what NV is meant to be here.

consensuses;[12] (d) conceptual frameworks or moral and political theories; (e) researcher's intuitions; (f) disciplinary or theoretical standards.[13]

I will exemplify how normative validation according to these *desiderata* and sources is practiced in poverty measurement. For now, I note that each source has its pros and cons. Still, if we accept democratic legitimacy as a pivotal principle in guiding normative judgments when designing a measure (Alexandrova & Fabian, 2022), then we must prioritize *a*, *b* and *c*.[14]

1.V. Objections to normative validity

To finish Part 1, I will discuss three objections to NV.

(1) NV is a type of content validity, therefore, redundant.

Well, when a notion is a type of a familiar one, it does not follow that we completely know, master, or pay due attention to it. While we shall see that NV just named something that already exists for a long time, many researchers still lack awareness of measurement design normative judgments. Also, NV deserves an articulation that fully establishes its consequences.

(2) NV goes against the value-free ideal.

Poverty is one of these concepts consensually taken as thick. Does this mean it goes against value-freedom? We may think that when measuring them, the burden of proof lies on the value-freedom enthusiast, as in the literature on measuring these concepts many authors claim that it is impossible or undesirable to accomplish value-free measurement (*e.g.*, Reiss, 2017; Chamberlain, 2021). Furthermore, let us consider the researchers dedicated to measuring poverty and inequality. Almost all of them take value-ladenness for granted and consider pursuing the value freedom ideal undesirable.

However, let us discuss an interesting possibility for a researcher that accepts the value freedom ideal *and* that poverty is a thick concept.[15] Suppose that Maria, an economist, wants to

[12] These are shared ideas in a political community, often translated into authoritative documents (*e.g.*, a constitution, the Universal Declaration of Human Rights) or international agreements (*e.g.*, the Sustainable Development Goals).
[13] For instance, economists' tendency of choosing preference satisfaction as an indicator for utility.
[14] Some measures mix different sources, like Bradshaw and Finch (2003) and the MPI-LA we shall discuss below (see also *2.II.a*).
[15] I am grateful to Anna Alexandrova for pointing this out to me. She also hypothesizes that such an attitude is widespread in scientists measuring basic things like poverty, inequality and deprivation. Further, she suggests that these "researchers who take on to study these phenomena have some ambition not to saddle science with the task of

measure the poverty of a country like Brazil. Suppose further that Maria discover through an opinion survey that the majority of Brazilian society held moral and political values about poverty different from her own. For instance, the majority of this society thinks that poverty's deprivation space is economic resources. Differently, Maria thinks that poverty's space are basic capabilities. What space must Maria take when constructing this society's poverty line? Well Maria can choose to build her measure not according to her values but of Brazilian society. She takes society's values as *constituents* of poverty but not as objects of her own evaluation or prescription. In the words of Sen, Maria's "poverty description will then *reflect* socially held value judgments rather than *be* value judgments themselves" (1980: 366, original emphasis). In fact, this is not just a theoretical possibility but it corresponds to an approach in poverty measurement called the consensual approach (Mack & Lansley, 1985). The normative judgments made during poverty measurement are thus deferred to society and explored through deliberative exercises, participatory methods, surveys or enduring consensuses.

(3) *NV adds to the undesirable proliferation of measures, undermining measurement's authority.*

Let us unpack this argument. Disagreements about values are endemic (in philosophy and also in morality and politics). Also, "the language of measurement connotes epistemic authority" (Mari *et al.*, 2022: xx), and this authority offers a common evidential ground that helps people solve this type of disagreement. Then, by welcoming NV, we risk weakening a powerful consensus-making tool. Well, many disagreements on measurement issues derive from normative disagreements that have not *yet been recognized* as such. NV either helps solve them or, if they persist, tells us *why* they exist. In fact, by requiring transparency, NV might reveal that what is considered common evidential ground is bad evidence because it is based on hidden normative disagreements. Moreover, there is a tendency to inflate the actual extent of normative disagreements. Enduring international consensus is the most significant evidence in favor of relevant normative consensus.

What we may call *normative awareness* has developed a research program for designing measures that can better reflect non-epistemic values. Attention to NV adds value theoretically,

effecting moral progress" (personal communication). The only important note is that the innovation of methodologies like the consensual approach is to strive for empirically looking out for society's value judgments and, more fundamentally, do so motivated by a democratic ideal.

politically, and empirically (Alkire *et al.*, 2015: 20). But it is, as we have seen above, *one* type of validity. In practice, designers can be justified in trading off considerations like the benefits of standardization, respect for previous knowledge and practice, and implementation issues at the expense of NV.[16]

PART 2: THE CASE OF POVERTY MEASURES

2.I. Normative validity from the standpoint of poverty measures

My answers to objections *2* and *3* above show how poverty measures peculiarly illustrate NV. In addition, poverty measures are insightful. It is productive to extend validity considerations to different disciplines from psychometrics, which was originally developed to measure subjective phenomena such as intelligence, happiness, or well-being – all taken as *latent variables*. However, in the best scenario, psychometrics only takes us halfway through. Even if these subjective elements have a role in measuring poverty, we must deal with aggregation issues that are, *prima face*, beyond the reach of psychology. These issues are the bread and butter of research programs like the capability and axiomatic approaches to poverty measurement, dedicated to devising measures that are more and more normatively valid – and here, RTM and RTM* are especially useful.

NV can contribute not only to the validity literature but also to a crucial internal dispute in poverty measurement – *i.e.*, between those researchers who privilege explanatory/predictive goals and those who privilege normative goals. It is fruitful to frame this as a dispute between privileging NV, a species of conceptually-driven validity, or privileging what I have called data-driven validities (convergent/divergent validity and nomological/construct validity).

But first, let us briefly check some normative judgments one faces when designing a poverty measure.

[16] A note about *robustness* tests, which check whether a normative-driven measure's results are not too sensitive to the choice of key parameters – some determined by normative choices. Robustness gives evidence that a measure's results are *reliable* and that different measures redundant, which helps establish a common ground. Although robustness and reliability are related to NV, there is no space to discuss this further.

2.II. Measuring poverty

There are many ways in which we can design a poverty measure. I will work here with Amartya Sen's (1976) two steps for designing a poverty measure: *identification* and *aggregation*. The former comprises finding a criterion to identify *individuals* in a given population who are in poverty.[17] Here, an individual may mean a person, but it may also mean households, families, communities, regions, countries, or even the whole world. Identifying poor individuals is an essential property for administrative measures of poverty: measures that serve as criteria for identifying those who can receive social benefits, like cash transfers.[18] Other measures concern poverty not of individuals but among *populations* (*e.g.*, whether poverty in a country increased compared to last year). For this, we need to sum up individuals in poverty.

As we check each step, I will show instances of how normative validation is applied to satisfy our two *desiderata* and justify their sources for normative judgments.

2.II.a. Identification: what are the concept and dimensions of poverty?

Poverty is a type of deprivation, and there are different types of deprivation related to poverty. The identification step involves defining what *thing* people in poverty are deprived of and at *what level* a person should be deprived of to be poor.

Our *characterization* of poverty guides the definition of what a person in poverty lacks. We can say that what defines each poverty concept presented to date is its *deprivation space*: the type of deprivation that a person in poverty faces. The most common concepts of poverty characterize it as a deprivation of: income, basic needs, basic resources, well-being, and basic capabilities.

Choosing one space may be a normative judgment for two reasons. It may entail a prescriptive judgment. There might be something in the *meaning* of the concept of poverty that tells us it is a type of unacceptable deprivation or something that every person has a stand on as a member of society. In the last sense, to choose the deprivation space, we should look at source *a* –

[17] Some measures do not identify individuals but just give us a general assessment of a population's poverty. Hence, Sen's account is a partial one.

[18] A notable example is the poverty measure of Brazil's conditional cash transfer, *Programa Bolsa Família*, which stipulates as poor those families whose monthly income *per capita* is below BRL 210,00 (*circa* 83 in PPP dollars). Note that if restricted to this goal, aggregation is unnecessary.

i.e., what political philosophy tells us about the currency of egalitarian justice. Second, as poverty is a *Ballung* concept (see *1.II*), it has many meanings. A normative reason may tell us which one we should choose to animate our measure. Moreover, in the last decades, there has been a growing consensus that poverty consists of not one but deprivations in several dimensions – *i.e.*, poverty is *multidimensional*.[19] In this case, characterizing poverty involves choosing a set of possible dimensions. We may regard this choice as an evaluative judgment. For instance, basing their choice on a mix of theoretical inputs, participatory and deliberative exercises, Wolff and De-Shalit (2007) propose the following dimensions: life, bodily health and integrity, imagination, control over environment, and socialization.

A second part of identification concerns defining the poverty *threshold*: where to draw the line that divides the poor from the non-poor. This question has occupied the minds of poverty researchers since the beginning of its scientific measurement. Some have claimed this is a matter of judgment, normative through and through. Others have tried to combine normative considerations with empirical criteria, like physiological needs. An exception is the British sociologist Peter Townsend, who did not deny that normative judgments end up playing a role in setting the poverty line but tried to minimize their influence. Hence, he endorsed the value-freedom ideal as a model of conduct that should guide scientific practice because, even if it is not possible to eliminate values from its practice, there are reasons that scientists should pursue this ideal. To do so, he privileged the power of measure to explain/predict people's behavior (1979: 46-7, 60).[20] More concretely, he hypothesized a discontinuity between the level of deprivation of the poor and the non-poor – there, we should draw the line. In our terms, Townsend aimed at a data-driven valid measure.[21]

2.II.b. Aggregation: which axioms?

The second step in measuring poverty concerns aggregating poor individuals or constructing a poverty *index* that "summarizes the information about poverty across society" (Alkire *et al.*, 2015:

[19] Even if the poverty deprivation space consists of one type of deprivation, this deprivation may be itself multidimensional (*e.g.*, capabilities are multidimensional).
[20] See note 10.
[21] Some reject drawing a poverty line and conceive poverty measures as special cases of social welfare functions, with a *continuum* between most and less deprived. However, these measures cannot count the poor or determine the level and distribution of poverty (Deaton, 1997: 140-44).

33). This abstract entity precisifies intuitive and imprecise ideas like "there's more poverty in Brazil than in Sweden" or "in 2022, poverty in Brazil returned to its level ten years ago." Indexes provide us with a single or composite value whose purpose is to determine information like the level and the dynamics of poverty in a population.

The most popular index is the *headcount ratio*, which consists of the ratio between the number of poor and the total population, providing us with the proportion, or *incidence*, of people in poverty. For simplicity, hereafter, we will consider a unidimensional poverty measure in the income space. From a normative point of view, the headcount ratio has quite known shortcomings. First, poverty varies in *depth* – *i.e.*, people in poverty may be closer or far away from the poverty line. According to the *monotonicity* principle, poverty in a society should rise if a person who is already poor loses more income, even if the poverty level remains the same. However, the headcount ratio is not sensitive about this. Worse, following the headcount ratio, the policy that would most reduce poverty would focus on those closest to the poverty line, thus overlooking the extremely poor.

Alternatively, the *poverty gap ratio* consists of the mean shortfall in income from the poverty line. Hence, it satisfies monotonicity (if someone who was already poor becomes poorer, the mean shortfall will increase) and is sensitive to how far, on average, people in poverty are below the poverty line. Still, poverty gap is insensitive to another important feature of poverty – *i.e.*, inequality among the poor. The same average gap in poverty can hide varying levels of inequality below the poverty line. Also, it is insensitive to income transfers among the poor.

The *weighted poverty gap* answers these issues by weighing individuals farther from the poverty line more heavily. Therefore, it satisfies the *Pigou-Dalton transfer principle*, which in the context of poverty measurement states that, even if poverty incidence and depth remain the same, a progressive income transfer between the poor – *i.e.*, from a poorer person to a less poor person – decreases general poverty.

The axiomatic approach provides a mathematical structure for systematically dealing with these normative principles. As Alkire *et al.* (2015: 51) say, this approach assesses poverty measures according to how its structure and response to change in its arguments satisfies certain axioms, some technical, like transitivity, and others normative, like monotonicity. In practice, satisfying normative principles means that, given data transformations (*e.g.*, someone who was already poor

becomes poorer), the measure values will behave according to these principles (*e.g.*, general poverty will increase).

Let us compare the axiomatic approach and traditional RTM. The main goal of the latter is to show, by proving representation theorems, that there is a proper meshing between how the relevant concept and its empirical structure are characterized and its formal representation in a numerical structure. By proving representation theorems, the former aims to show that there is a proper meshing between how the normative principle is satisfied given the related data transformations and its formal representation in a numerical structure. In the axiomatic approach, the axioms are not descriptive *dictums* (*e.g.*, about how people in poverty behave); they are normative *dictums* (*e.g.*, about how the poverty measure's results should behave given changes in the population).[22]

For NV, this approach satisfies the transparency *desiderata* the most. Also, it provides a formal, deductive argument that all three measurement elements mesh properly: the representation – index –adequately captures one or more normative principles consistent with the concept of poverty – characterization – because, given data transformations, the values of the measure – procedures' output – behave under these principles. However, it is not normatively foolproof. A measure may lack NV because we take one of its axioms as inconsistent with the conceptual framework that animates our measure (source *a*), with people's view about the matter (source *b*), or with authoritative documents (source *c*).[23] Unfortunately, the common practice in the literature is choosing axioms based on a researcher's intuition or disciplinary standards (respectively, sources *d* and *e*). Beck *et al.* (2020) show how normative philosophy, combined with philosophy of science, can assess the justifiability of sets of axioms or question who chooses them.

[22] *Cf.* Angner's (2011) account of measurement approaches in welfare economics. While certainly useful, his account does not pay due attention to the application of axiomatics for normative purposes. Interestingly, against orthodox economics, an empiricist could say that axiomatization is a more appropriate tool for prescription, not description.

[23] Chamberlain (2021) claims that the global Multidimensional Poverty Index (MPI) is inconsistent. One of the MPI's axioms is that it should provide total orderings of people in poverty. To achieve this, different dimensions need to be traded off (*e.g.*, trading off health for the sake of education). This, however, contradicts the MPI's conceptual framework, the capability approach. This approach forbids compensating between different capabilities, which means that only partial orderings are possible. Nevertheless, total ordering may be necessary in certain contexts, like allocating resources for a region with the most poverty. While measures like the MPI have less NV according to the capability approach than alternatives with partial rankings, according to other normative criteria (*e.g.*, efficiency), they may have more NV.

2.III. Normativity or explanation? The MPI for Latin America

I will finish the paper showing how normative validation practices in poverty measurement can conflict with convergent/discriminant and nomological/construct validation. Comparing what I have called measures that privilege normative-driven validity and those that privilege data-driven validities can help us explain old poverty measurement disputes that still echo today.

One of the most important motivations for measuring poverty is to help design *interventions* that fight poverty. Interventions ask for causal or predictive analyses, and the main goal of data-driven poverty measures is to be adequate for them. These measures comprise a *latent variable representing the systematized concept* and a *statistical model* connecting the latent variable with observed variables. Here, convergent/divergent and nomological/construction validation determine whether a poverty measure's results are, respectively, effects of the concept or compatible with expectations derived from a well-established causal hypothesis about poverty.

Borrowing from Greene's (2020) discussion of James Woodward's recommendations, a poverty measure should have some properties to feature in a causal model. Its variable needs to be homogeneous, unambiguous, and simpler. These properties constitute guidelines for a concept's characterization. For instance, to measure multidimensional poverty, we must select a few or one of its meanings and reduce its dimensions to model it appropriately. An important tool for reducing dimensions is precisely convergent/discriminant validation, as one of its goals is to combine parsimony while retaining most of the multidimensional information. The ultimate goal is to test the model, which translates into performing construct validation. So, we choose poverty dimensions according to their descriptive and explanatory/predictive power.

Compare this with the case of our measure being normative-driven. We should choose its dimensions not according to statistical but normative criteria. For instance, if our measure follows an axiomatic approach and is informed by a conceptual framework like the capability approach (Comin *et al.*, 2008), we should choose dimensions corresponding to basic capabilities. Capabilities have an intrinsic value, and a measure that does not capture these capabilities would not be normatively valid. It does not make sense to drop a dimension because it has less explanatory/predictive power.

A recent debate (Catalán & Gordon, 2020; Santos & Villatoro, 2020; Gordon & Catalán, 2020) illustrates these differences nicely. Hector Catalán and David Gordon (C&G) promote an

analysis of, *inter alia*, the validity of the Multidimensional Poverty Index for Latin America (MPI-LA), designed by Maria Santos and Pablo Villatoro (S&V). It is based on the axiomatic and the capabilities approaches. It has 13 indicators of 5 dimensions (housing, basic services, living standard, education, and employment and social protection) taken to constitute poverty. Here, I only consider features of the MPI-LA related to validity.[24]

C&G analysis build upon Townsend's approach to poverty measurement (Catalán & Gordon, 2020: 1763-64) and psychometric's approach to validity (1777-78). Accordingly, a poverty measure should capture manifestations of an underlying phenomenon, poverty. For this, the measure is designed around a latent variable that, in turn, should structure and be part of a higher-order construct, poverty. For MPI-LA to satisfy these *desiderata*, it must be convergent/discriminate valid – *i.e.*, its dimensions must display explanatory/predictive power. This would mean that the MPI-LA can capture a latent variable and that this variable fits in an explanatory/predictive model that specifies its relationship with other variables. These other variables are poverty *indicators* that, in turn, should capture the 5 dimensions. Satisfying these *desiderata*, the MPI-LA could be assessed according to construct validity, which means testing if the indicators are caused by multidimensional poverty and no other phenomena.

C&G claim that the MPI-LA does not pass these tests. It is statistically inconsistent, with low evidence that these indicators can capture a higher-order construct – *i.e.*, poverty. Between the lines, they are saying, first, that the axiomatic approach is a bad validation tool, as it advances a measure whose structure does not provide a statistical model of poverty and, therefore, cannot capture a latent variable. Second, the capability approach criterion for selecting the dimensions is arbitrary – what the researcher considers "bad" (adaptation from Gordon, 1995: 39, cited in Gordon & Villatoro, 2020: 1792). Third, it does not allow us to identify a parsimonious and "optimum sub-set of deprivation [poverty dimensions] indicators" (1792). In other words, the MPI-LA is invalid according to data-driven validities' standards.

In their answer, S&V argued that C&G assessed the MPI-LA according to their preferred standards, which the measure did not intend to meet. They emphasized that the axiomatic approach makes the normative judgments that animate the MPI-LA transparent. Also, in choosing their dimensions, they tried to reflect normative priorities of Latin American countries (they mix sources

[24] For instance, I do not discuss its reliability (see note 16).

a and *c*, *1.IV*). More importantly, S&V say that the MPI-LA is not a statistical model of poverty that aims at a latent variable. It is, in their words, only a measurement instrument for capturing a characterization of poverty that clusters different deprivations in the same thematic domain. For S&V, lacking a statistical model does not hinder MPI-LA's validity because:

- its characterization is theoretically and conceptually based;
- its representation follows an axiomatic structure;
- if there are data transformations, its results behave under normative principles.

Under the terms presented here, S&V claim that MPI-LA is valid according to NV's standards. They claim that the MPI-LA is focused on reflecting what people in Latin America take as their normative priorities. What S&V identified as these priorities are made clear because of the measure's axiomatic structure. They may have identified them wrongly, but their measure enables us to pinpoint the locus of their mistake. Also, statistical tools did not (and must not) determine MPI-LA dimensions as these represent basic capabilities, which are seen as endowed with intrinsic value by their approach.[25]

And what is my diagnosis? As a takeaway, I think that if a researcher is assessing the validity *simpliciter* of any poverty measure, she must at least *consider* the existence of NV. It may be the case that, according to the purposes of the measure, NV may weigh less than data-driven validities. Still, this differs from denying any place for NV or normative-driven measures.

C&G would be right *only if* explanation is the sole or paramount purpose a poverty measure should satisfy, as a measure of explanatory/predictive power would help satisfy it. This is how they could claim that their statistical approach is the gold standard for measuring poverty. For sure, explanation is a valuable purpose, but it is not the only one (*cf. 1.I.b* validity's fourth feature). Statistics may even hamper making justice to the normative features of poverty and suggest a measure that opposes commonsensical understandings of it. Characterizations and poverty measure's results based on statistical approaches are often revisionary – *i.e.*, identifying as poor individuals who are usually considered non-poor. Such a revision may go against a measure purpose.

[25] Nevertheless, the MPI-LA provides total orderings, an inconsistency with the capability approach (see note 23).

C&G's stance is radical. Most poverty researchers believe no single measure or method is the best *simpliciter*. Different measures might capture different things about the concept of poverty (Cowell, 2016: 26; Lister, 2021) and might be designed for different purposes and contexts (Sen, 2018: 29-30; Atkinson 2019: 28). "But give me a reason why explanation/prediction should not be privileged. Even if not sufficient for fighting poverty, they are certainly necessary," one may say. I submit. However, we should not confine normative considerations or goals to our current and provisional explanatory and predictive tools. We may develop improved tools where measures with high normative validity fit well. If this never occurs, we will still have a knowledge of irreducible value that normatively valid poverty measures can provide the most: the state of deprivation in our societies, given what we value the most.

REFERENCES

Adcock, R.; Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review, 95*(3), 529-46.

Alexandrova, A. (2017). *A philosophy for the science of well-being*. Oxford: Oxford University Press.

Alexandrova, A.; Fabian, M. (2022). Democratising Measurement: or Why Thick Concepts Call for Coproduction. *European Journal of Philosophy of Science,* 12(7).

Alkire, S.; Foster, J.; Seth, S.; Santos, M. E.; Roche, J. M.; Ballon, P. (2015). *Multidimensional poverty measurement and analysis: A counting approach*. Oxford: Oxford University Press.

Anderson, E. (2002). Situated knowledge and the interplay of value judgments and evidence in scientific inquiry. *In* Gärdenfors; Woleński; Kijania-Placek (eds.). *The Scope of Logic, Methodology and Philosophy of Science*. Springer, 497-517.

Atkinson, A. B. (2019). *Measuring Poverty Around the World*. Princeton-NJ: Princeton University Press.

Babbie, E. R. (2014). *The Practice of Social Research (14th edition)*. Boston-MA: Cengage Learning.

Beck, V.; Hahn, H.; Lepenies, R. (eds.) (2020). *Dimensions of Poverty: Measurement, Epistemic Injustices, Activism*. Cham: Springer.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley and Sons.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Bradshaw, J.; Finch, N. (2003). Overlaps in dimensions of poverty. *Journal of Social Policy,* 32(4): 513-525.

Cartwright, N.; Bradburn, N. (2011) A theory of measurement. *In* National Research Council (org.). *The Importance of Common Metrics for Advancing Social Science Theory and Research: A Workshop Summary*. Washington-DC: The National Academies Press, 53-56.

Catalán, H. N.; Gordon, D. (2020). The importance of reliability and construct validity in multidimensional poverty measurement. *The Journal of Development Studies,* 56(9), 1763-83.

Cowell, F. (2016). Inequality and Poverty Measures. *In* Adler & Fleurbaey (eds.). *The Oxford Handbook of Well-Being and Public Policy*. Oxford: Oxford University Press.

Chamberlain, T. D. (2021). *The Capabilities Approach to Well-being: Characterizing Capabilities and Measuring Them*. Ph.D. Thesis. University of California, San Diego.

Comin, F.; Qizilbash; M., Alkire, S. (eds.) (2008). *The Capability Approach: Concepts, measures and applications*. Cambridge: Cambridge University Press.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington-DC: World Bank.

Deeming, D. (2020). *Minimum Income Standards and Reference Budgets: International and Comparative Policy Perspectives.* Bristol: Policy Press.

Elliott, K. C. (2017). *A Tapestry of Values: An Introduction to Values in Science*. Oxford: Oxford University Press.

Greene, C. (2020). Nomadic Concepts, Variable Choice, and the Social Sciences. *Philosophy of the Social Sciences*, 50(3), 3-22.

Goldthorpe, J. H. (2015). *Sociology as a population science.* Cambridge: Cambridge University Press.

Kahneman, D.; Wakker, P. P.; Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, 112(2), 375-405.

Krantz, D.; Luce, R.; Suppes, P.; Tversky, A. (1971). *Foundations of Measurement, Vol. I: Additive and polynomial representations*. Cambridge-MA: Academic Press.

Larroulet Philippi, C. (2020). Valid for What? On the Very Idea of Unconditional Validity. *Philosophy of the Social Sciences*, 51(2), 151-75.

Lister, R. (2021). *Poverty (2nd edition*). Cambridge: Polity.

Mack, J.; Lansley, S. (1985). *Poor Britain*. London: Allen and Unwin.

MacPhail, F. (1998). Moving beyond Statistical Validity in Economics. *Social Indicators Research*, 45(1/3), 119-49.

Mari, L.; Wilson, M.; Maul, A. (2021). *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*. Springer.

Messick, S. (1989). Validity. *In* Linn (ed.). *Educational measurement (3rd edition)*. London: Macmillan/American Council on Education.

Gordon, D.; Catalán, H. N. (2020). Reply to Santos and Colleagues 'The Importance of Reliability in the Multidimensional Poverty Index for Latin America (MPI-LA).' *The Journal of Development Studies*, 56(9), 1790-94.

Piachaud, D. (1981). Peter Townsend and the Holy Grail. *New Society*, 10 September, 419–21.

Reiss, J. (2017). Fact-value entanglement in positive economics. *Journal of Economic Methodology*, 24(2), 134-49.

—, (2022). Measurement and Value Judgments. *In* Heilmann & Reiss (eds.). *The Routledge Handbook of Philosophy of Economics*. New York: Routledge.

Santos, M. E.; Villatoro, P. (2020). Response: The Importance of Reliability in the Multidimensional Poverty Index for Latin America (MPI-LA). *The Journal of Development Studies*, 56(9), 1784-89.

Sen, A. K. (1976). Poverty: An Ordinal Approach to Measurement. *Econometrica*, 44(2), 219-31.

—, (2018). *Collective Choice and Social Welfare: An Expanded Edition*. Cambridge-MA: Harvard University Press.

Zeller, R. A.; Carmines, E. G. (1980). *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge: Cambridge University Press.