Studies in History and Philosophy of Science 113 (2025) 88–97

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Measurement requires compromises: the case of economic inequality

Alessandra Basso [a,*], Anna Alexandrova [b]

[a] London School of Economics, Department Philosophy Logic and Scientific Method, United Kingdom
[b] University of Cambridge, Department of History and Philosophy of Science, United Kingdom

ARTICLE INFO

Keywords:
Measurement
Human sciences
Philosophy of measurement
Quality of measurement
Validity
Social indicators

ABSTRACT

We examine considerations that enter into design and evaluation of measures in social science, categorizing them into four drivers: epistemic, ethical, pragmatic, and metrological. We call them drivers to highlight their role in guiding researchers' decisions without determining them. Through an analysis of the World Inequality Report 2022, we reveal tensions among these drivers, illustrating the complex interplay between the various demands a measure must satisfy. Our analysis highlights the need for case-by-case compromises to address these tensions, as optimizing one driver often comes at the expense of another. We explore the extent to which these compromises shape measurement practice and the principles that guide researchers in balancing them. While existing literature on measurement assumes that tensions can be resolved with good practice and use, we argue that developing a good measure requires balancing multiple demands, recognising that it might be impossible to meet all of them simultaneously.

1. Introduction

Measurement has more than one aim: it gives empirical content to theoretical parameters, enables comparisons, facilitates accumulation of knowledge across time and contexts, informs decision-making and evaluation, and provides an empirical lens for grasping nebulous yet socially relevant phenomena, thus keeping those responsible for them accountable. Users and producers of metrics at times acknowledge tensions between these aims. Still, a generalized view is that a good measurement achieves all its intended aims. This view is implicit in various accounts of measurement. Mari et al. (2023)'s *Measurement Across the Sciences* presents a framework for understanding measurement across the physical and human sciences. In their view, a good measurement is objective and intersubjective, that is, it conveys information about the measured property and remains interpretable across different observers and contexts. While the authors acknowledge tensions between these qualities (p.83), the framework progresses step by step towards establishing a public value for the measured property, thereby securing both desired qualities. The framework leaves no room for achieving these qualities only partially. In another cutting-edge contribution, McClimans' (2024) *Patient-Centred Measurement* explores tensions between the goal of standardisation and the need to represent the diverse, evolving perspectives of patients. Despite the challenges these tensions pose, the book remains optimistic that proper development, interpretation, and use can resolve them.

Looking at an ambitious example of measurement in the social sciences, *The World Inequality Report 2022* (Chancel et al., 2022), we argue that it is often impossible to design measurements that fulfil all their intended purposes. This is because there are tensions between different aims and needs, such that achieving one aim may prevent the satisfaction of another.

The idea that measurement requires compromises has been voiced before. Scientists sometimes acknowledge making compromises when discussing the epistemic costs of pragmatic choices, like those made during data collection or measurement implementation (e.g.,van Drimmelen et al., 2024; Ioannidis, 2005; Simmons, Nelson, and Simonsohn, 2011). Historians and philosophers of science emphasize that development and evaluation of measures requires prioritizing among different aims, only some of which are epistemic (Boumans, 2015; Bradburn et al., 2017; Cartwright and Runhardt, 2014; de Vet et al., 2011). According to Cartwright and Runhardt (2014, p. 276), trade-offs are so pervasive in scientific practice, especially in the social sciences, that addressing them is a central problem that scientists continuously confront.

This article is part of a special issue entitled: Measuring the Human published in Studies in History and Philosophy of Science.

* Corresponding author. London School of Economics and Political Science, Department of Philosophy Logic and Scientific Method, The Lakatos Building, Houghton St, London, WC2A 2AE, United Kingdom.
E-mail address: a.basso@lse.ac.uk (A. Basso).

https://doi.org/10.1016/j.shpsa.2025.08.005
Received 26 June 2024; Received in revised form 15 June 2025; Accepted 9 August 2025
Available online 8 September 2025

However, the extent and modality with which these compromises shape measurement practice have not been recognized, nor documented in detail. Trade-offs emerge from specific decisions during development and validation, and discussions on the topic have so far been highly contextual, often tied to the contingencies of individual cases. Trade-offs have been identified between different epistemic values, such as precision, calibration, and standardisation (Boumans, 2015, p. 84), or between epistemic aims and other considerations, including ethical and pragmatic needs (Cartwright and Runhardt, 2014). However, there has been little effort to systematize the types of tensions that arise in measurement practice. We address this question by asking whether the trade-offs are expressions of broader tensions between multiple aims of measurement, rather than being confined to individual cases. To do this, we explore the range of considerations that guide researchers in the development, evaluation, and use of a measure and identify four main drivers of their choices. Our analysis suggests that the trade-offs arise from broad tensions between the drivers, which span various choices.

This systematization also enables us to examine the principles that guide researchers in addressing these tensions, a topic that is largely unexplored in the philosophy of measurement. Boumans (2015) argues that no epistemic principle can dictate how to resolve the tensions; instead, addressing them requires expert judgment. However, he does not address the considerations that might guide this judgment. Some insights can be found in the values in science literature, where the recognition that scientific practice can serve a variety of aims, both epistemic and non-epistemic, is often coupled with the claim that value influences are acceptable as long as they accord with the epistemic goals of inquiry (Potochnik, 2020; Rosales and Morton, 2019). On this view, moral values play a secondary role, either closing the gap between evidence and theories (Longino, 1990, 2002) or managing the risks of error associated with uncertainty (Douglas, 2009).[1] This perspective positions ethical considerations as peripheral, invoked only when epistemic ones are insufficient (Elliott, 2011; Steel, 2010). In contrast to this view, we find that in measurement there is no lexical order of priority between different aims, nor a general rule for determining which one should trump the others. This suggests that different sets of principles may be applied depending on the case.

In the rest of the article, we articulate how the need to address tensions and trade-offs shapes the *World Inequality Report 2022* (hereafter WIR). After a brief primer on measurement of inequality in Section 2, Section 3 identifies four drivers that can in principle guide the researchers' choices: epistemic, ethical, pragmatic, and metrological. We show that all four can play a role and that each one opens up a variety of options, none of which is superior to the others. In Section 4, we examine an influential approach in the measurement of economic inequality developed in the Paris School of Economics and embodied in the WIR. Specific versions of the four drivers emerge when analysing the Report's key choices and their justifications. Based on this case study, in Section 5 we highlight three broad tensions between the drivers, reflecting the complex interactions between the diverse needs and purposes of social indicators. The first tension arises between epistemic and ethical drivers, the second between metrological and epistemic drivers, and the third involves all four drivers in a trade-off between fruitfulness and visibility of measurement. We then look at how these tensions are addressed in the Report to see if there are consistent principles guiding the economists' choices, or signs of relations of priority between the different aims. Though grounded in the case study, this analysis will be useful more generally in two ways: first, by offering a vocabulary for scientists and philosophers of science to articulate these compromises, and second, by serving as a step toward developing sets of principles to navigate these decisions.

[1] Another view posits that scientific assessment requires the *joint* satisfaction of epistemic and non-epistemic values (Lusk and Elliott, 2022). This approach resonates with the idea that a good measurement is one that fulfils all its aims.

2. Measurement of inequality: a primer

To study inequality, economists measure the degree of concentration of economic resources across a population, and use these measurements to make comparisons across regions and monitor trends over time. Inequality is quantified by summarizing statistics that associate any distribution of resources with a number that captures the degree of concentration, thereby ensuring a complete ordering of distributions in terms of inequality.

Inequality indicators are developed within an axiomatic approach, which starts off by identifying a set of axioms that the indicators must respect. These axioms, commonly called principles, describe properties of inequality that are considered self-evident and provide the basis for constructing an indicator. For example, the anonymity principle states that inequality is indifferent to the identifying features of the individuals in the population, so that if two individuals exchange their resources, the population inequality remains the same. The population principle, instead, states that inequality is independent of the size of the population, so that if the population is multiplied x times, inequality remains the same. The transfer principle states that inequality decreases when there is a transfer of resources from a rich to a poor that does not change the order between them. The number and kind of axioms vary depending on the features of inequality the researchers want to capture. The higher the number of axioms, the fewer indicators are able to satisfy all of them (Costa and Pérez-Duarte, 2019).

The axioms alone, however, do not cover all the choices researchers face when constructing and evaluating an indicator of inequality. Economists discuss about a variety of other choices, considering the implications of various alternatives and evaluating their appropriateness from multiple perspectives (see e.g., Atkinson, 1983; Bojer, 2003; Salverda, Nolan, and Smeeding, 2009; Sen, 1997a).[2] For example, in abstract terms inequality is a property of the distribution of resources across a population, but to construct an indicator researchers must choose which resource to focus on, the relevant population, and the statistical unit. Economic inequality is usually defined in terms of income or wealth, but there are several alternative options including consumption, utility, or capabilities – and important distinctions between inequality of outcomes and inequality of opportunities. With respect to the population and statistical unit, common options include measuring inequality between households within a country, global inequality between countries, or global inequality among individuals regardless of where they live. Constructing an inequality indicator requires making several additional choices, for instance about how to deal with specific forms of income, like dividends, social benefits, deductions, and non-monetary incomes. Yet other choices regard equivalence scales, summarizing statistics, and data sources. Alternative equivalence scales have been proposed to compare households of different sizes and compositions. Data are usually drawn from income surveys and national censuses, but one can also consider other sources like tax records and national accounts.

Based on these choices, researchers construct the distribution of the resource of interest across the relevant population and then use statistical tools to summarize the inequality of the distribution into a single number. Alternative summarizing statistics are available, like the Gini, Theil, and Atkinson indexes, or percentiles ratios like the P90/P10.[3]

These choices can significantly influence the study results. Not only, as one can expect, are there large differences in inequality depending on the specific resource one focusses on – wealth, income, or consumption

[2] Textbooks and methodological contributions summarize these discussions considering alternative options for each aspect of the indicator that requires specification (see e.g., Afonso et al., 2015; Baldini and Toso, 2009; Jenkins et al., 2009; McGregor et al., 2019; Piketty, 2015; Ray, 1998).

[3] The P90/P10 ratio compares the resources at the 90th percentiles to the ones at the 10th percentile.

for example. But also, other choices have significant impacts on the conclusions drawn from inequality indicators. For instance, it has been shown that relying on different indexes can lead to inverse ranking, other things being equal (Jenkins, 2024). Moreover, inequality trends can appear different depending on the statistical units or equivalizing scales employed (Atkinson and Brandolini, 2009; Jenkins, 2015). The definition of the population of interest also has a significant impact on the measurement outcomes: recent trends in global inequality can appear different depending on whether one looks at inequality between countries or inequality between individuals regardless of where they live (Lakner and Milanovic, 2016). As a consequence, the quality of the indicator and its authority in arbitrating decision-making processes depend on how these choices are made and justified (Alexandrova, 2017; Cartwright and Runhardt, 2014; Hand, 2004; Reiss, 2008). This is evidenced by recent and ongoing debates about the current level and trends of global inequality, where scholars challenge each other's findings by questioning the specific choices made to construct the indicators (e.g., Auten and Splinter, 2023; Lakner and Milanovic, 2016).

3. Four drivers in the measurement of economic inequality

We are now in the position to organise the range of considerations that make a difference to the choice of indicators into distinct analytic categories. The economic literature often mixes different types of consideration while we seek to distinguish them according to the values they appeal to. We call them drivers to underline the fact that they serve as general motivations for methodological choices without determining these choices fully. We divide the drivers into epistemic, ethical, pragmatic, and metrological. These adjectives name the general value that each driver pushes: the epistemic driver gathers considerations that all roughly have to do with pursuit of knowledge, the ethical driver encompasses reasons that have to do with what is good and right in the moral sense, the pragmatic driver picks out considerations of practicality, and the metrological driver is all about values of quantification. Each driver is an umbrella term for a group of more specific reasons animated by distinct values. We acknowledge that these drivers are broad and heterogeneous, so our suggestion should not be seen as a neat classification of mutually exclusive categories. Moreover, some considerations are impure, so to speak, in that they mix two or more drivers together. Nevertheless, we find the idea of a driver and the four-way division helpful for bringing some interpretative order to the complex process by which an indicator is developed and assessed.

The next four subsections unpack each driver, connecting them to the literature in philosophy of measurement and showing where in the economic discussions these drivers are located. Our focus will be in highlighting the most salient aspects of these drivers for measurement practice.[4]

3.1. The epistemic driver

The epistemic driver is a collection of considerations that refer to those virtues of an indicator that ensure it is properly connected to the existing knowledge of the phenomenon that the indicator represents. These considerations include a theoretically grounded definition of the phenomenon, causal or correlational evidence that the indicator has a stable link with the phenomenon, and various markers of precision, accuracy, and reliability of this link. This driver is analogous to what philosophers often describe as epistemic values, but in our case the focus is on measurement practice, hence the emphasis on effective coordination between theoretical concepts and their empirical content.[5]

In the literature on philosophy of measurement, the epistemic driver is reflected in accounts of the nature of validity, as well as discussions of how scientists overcome uncertainty in the process of validation of measures.[6] To test whether the indicator tracks the phenomenon, scientists are supposed to formulate and confirm causal claims about the relation between the measure and the measurand (though some disagree). On some accounts, this is accomplished with axioms that enable proof of representation theorems relating the observations and the concept (Chang and Cartwright, 2008; Krantz et al., 1971; Vessonen, 2021b).[7] To capture the challenge of figuring out what's a good measure of a phenomenon that is itself unclearly understood, philosophers made use of notions like epistemic iteration (Chang, 2004), co-evolution of measurement and theory (van Fraassen 2008), and model-based measurement (Tal, 2017). In the more applied literature in the sciences, the epistemic driver is explored in the various conceptions of validity (Lissitz, 2009; Zeller and Carmines, 1980) and uncertainty (JCGM, 2020; Giordani and Mari, 2013; Mari et al., 2023).

In case of inequality, the epistemic driver invites an evaluation of how well the indicator captures the relevant concept of inequality.[8] For example, when choosing how to define inequality, researchers are prompted to consider which resource is best suited for capturing the actual endowments of economic resources, independently of temporal fluctuations and distribution within micro-social organizations. An often-made remark is that income varies significantly over a lifetime, making the measurement of inequality at any specific time relevantly dependent on these fluctuations (e.g., Lakner and Milanovic, 2016; McGregor et al., 2019). This is undesirable because people can sometimes compensate for income fluctuations with savings and wealth, which means that income may not accurately reflect a person's or household's economic circumstance. If a household has an income gap, a measure of income inequality would place it at the bottom of the distribution, despite the fact that this household may be able to use savings and wealth to smooth out its financial situation.

Researchers can mitigate this worry by focussing on consumption, which instead is smoothed over a person's lifetime. However, consumption has other disadvantages because it depends on endogenous decisions: some people might live a more parsimonious life than their circumstances would allow, or on the contrary, consume more than they can afford by relying on credit. Therefore, consumption is not a 'clean' measure of people's circumstances either, though for different reasons. Neither of these options fully captures the complexity of people's economic circumstances, but the concerns they raise may be more relevant to certain purposes than others. An examination of economic inequality may consider multiple resources for a broader perspective; nevertheless, these epistemic considerations can help researchers decide which

[4] The idea that scientific theorizing can be assessed based on a range of theoretical virtues is well-established in the philosophy of science, along with the recognition that these virtues can sometimes conflict. Philosophers of science are well equipped to analyse compromises between competing epistemic aims (Kuhn, 1981; Levins, 1966) and, more recently, between epistemic and value-driven considerations (Douglas, 2009; Longino, 1990). Our goal here is to build on this work and make these connections more explicit in the context of measurement practice.

[5] In Mari et al. (2023, p. 259), these considerations are captured by the quality of object-relatedness (or objectivity), which they define as the extent to which the information provided by a measurement concerns the intended measurand and nothing else.

[6] For a discussion of the nature of validity, see e.g., Borsboom et al. (2004). An account of the process of validation can be found, for example, in Bradburn et al. (2017). Tal (2016) provides an account in which idealized models (used in the evaluation and management of measurement uncertainty) mediate between quantity concepts and their realization.

[7] Insofar as representational theorems establish the type of scale that can be used for a given measurement, they also respond to the metrological driver.

[8] In some cases, a measurement reflects a collection of attributes, requiring the researcher to address the problem of how to aggregate them. However, in the case of economic inequality, economists typically focus on the monetary value of resources like income and wealth, treating inequality as a one-dimensional parameter.

resource, or set of resources, to focus on.

Epistemic considerations bear on other choices too. For example, they can contribute to justify the choice of statistical unit (e.g., household or individual) and the equivalence scale used to compare households of various sizes. Methodological discussions emphasize that measuring individual inequality can create distortions because it does not reflect the social aggregation and sharing of resources within households (Coulter et al., 1992). On the other hand, household inequality raises other kinds of worry because households come in different sizes and compositions, and we lack knowledge about internal distribution. Neither of these options is clearly superior, but they raise different kinds of concern and therefore one might be preferable over the others depending on the intended purpose and context of analysis (e.g., different sources of error can be best controlled in distinct contexts).

3.2. Ethical

A second type of consideration is the ethical driver. These are considerations that affect the choice of measures by appeal to some moral reason having to do with justice, harm, wellbeing, rights. For example, researchers might invoke such reasons when they pick an indicator that explicitly highlights an injustice (or progress) in the most effective way. Or they might use value-based considerations to select the sensitivity of a metric or even its content.

There is no well-established literature highlighting the ethical driver. However, it is implicit in several traditions. The French movement of statactivism is dedicated to using statistics to effect activist goals typically by disclosing official statistics that reveal hidden costs of governance and regulation (Bruno et al., 2014).[9] American economists Anne Case and Angus Deaton use this sort of reasoning to invent a measure of 'deaths of despair', namely the premature deaths in deprived regions of USA due to opioid addition and suicide (Case and Deaton, 2020). Normative considerations enter the very definition of concepts and therefore the choice of metrics in cases where the object of study is picked out by a 'thick concept', that is, concepts that involve a descriptive and a moral dimension, like wellbeing, quality of life, poverty, inflation etc. (Alexandrova and Fabian, 2022; Angner, 2013; Cartwright and Runhardt, 2014; McClimans, 2010, 2024, Thoma, 2024).

Without trying to be exhaustive, we can unpack several ways in which ethical considerations can influence measurement. Ethical motivations might guide the selection of measures that better capture phenomena of ethical significance (Alexandrova and Fabian, 2022; Anderson, 2004). They can shape the background assumptions about what evidence is relevant for measuring a particular property (Longino, 1990; Peschard & van Fraassen, 2014). Research ethics may impose constraints to ensure ethical standards, such as protecting privacy, are upheld. Concerns about the social consequences of quantification might lead to choosing measures that promote morally desirable outcomes (Rodriguez Duque et al., 2024). Additionally, ethical considerations can affect decisions involving inductive risk, favouring designs that minimize the likelihood of more serious errors (Douglas, 2009).

In case of inequality, economists remark that ethical considerations are unavoidable when choosing how to measure inequality. Ranking levels of inequality is never purely statistical because it requires making judgments about how to weight resources at different points of the frequency distribution and which properties of the distribution are the focus of analysis (see e.g., Baldini and Toso, 2009; McGregor et al., 2019; Piketty, 2015). Any way of comparing different distributions in terms of their inequality must rely on a specific way of weighing resources in different parts of the distribution. For example, because of how it is constructed, the Gini index gives greater weight to changes around the middle of the distribution compared to transfers at top and bottom tails. This is because its sensitivity depends not on the size of the income levels but on the number of people situated between them (Sen, 1997a). This design has ethical implications: transfers within densely populated parts of the distribution influence the index more than transfers at the bottom tail, where fewer individuals are present but the needs are often more pressing. In this sense, ethical judgments are embedded in the weighting mechanism of any index. Therefore, when researchers choose a statistical tool to measure inequality, they implicitly make ethical judgements about the relative importance of resource transfers in different parts of the distribution. This illustrates what Schroeder (2019) describes as values fixed within the measurement (as opposed to being adjustable according to the user's needs). The weighting mechanism carries ethical implications and cannot be modified because it depends on the welfare function that is implied by different statistics.

The Atkinson index breaks with this tradition by embedding values in a way that makes them more user-accessible. In this index, the welfare weighting is controlled by a variable called 'inequality aversion', which researchers can adjust. By modifying this variable, they can change the sensitivity of the index to transfers at different points in the distribution and test how this affects the measurement outcomes (Sen, 1997a). This approach makes the value-laden choice explicit and encourages researchers to provide ethical justifications for their chosen value of inequality aversion.

Ethical considerations can shape other choices too, such as the definition of the population of interest. Researchers define the population under study based on the specific type of inequality that is morally significant for the purposes of the study (Baldini and Toso, 2009; McGregor et al., 2019). For example, one can choose to study inequality between countries because countries are actors in international diplomacy and their motivations for addressing global issues are determined by their position in a between-country distribution. Inequality between countries, for instance, is one of the primary arguments in favour of international aid and cooperation. Alternatively, researchers can choose to study global inequality among individuals regardless of where they live, because ultimately what matters is the situation of individual people, which is masked in broad groupings.

3.3. Pragmatic

The pragmatic driver refers to considerations arising from practical needs of measurement practice, such as resource constraints and ease of use. Datasets cost a lot of money because to assemble them requires time, human resources, and often material basis. Everything from negotiating access to existing data, to gathering your own, to processing, and storing them, is intensive in terms of labour and resources. So it is unsurprising that measures are often judged by appeal to values of feasibility, convenience, and tractability, and other practical considerations.

Despite the centrality of the pragmatic driver in the lived experience of scientists, these considerations are often regarded as playing only a secondary role. According to Hand (2004), pragmatic choices are needed for resolving aspects of measurement that are epistemically unforced, like selecting a unit of measurement or coming up with a precise definition for a loosely specified concept when relevant epistemic considerations are absent. On his account, whether these choices involve adhering to conventions or devising ad hoc solutions, they are made with the hope that their impact on the study's conclusions will be minimal. Ideally, researchers should restrict their conclusions to claims that are invariant to changes in pragmatic choices. However, achieving this ideal is often impossible, and therefore pragmatic considerations can have a significant influence on measurement processes.

The practical aspects of building and using indicators, like the process of data gathering and the employment of indicators in conjunction with other parts of science, are readily visible in the discussions on measurement of inequality. For example, economists remark that

[9] Morgan and Bach (2018) provide an extensive case study showing how measurements can prompt policy actions by serving as means for accountability and evaluation, as well as public instruments of persuasion.

General Entropy indicators (a class of indexes inspired by information theory) have the attractive property of being perfectly decomposable (e.g., Sen, 1997a). These indicators, such as the Theil index, allow to compare inequality between and within different subgroups of the population. This is because they can be obtained as a weighted sum of the values the indicator takes *within* different subgroups of the population, plus a term summarizing the inequality *between* the subgroups. For this reason, General Entropy indicators are preferable when studying the level and evolution of inequality within subgroups of the population, such as social groups or regional areas. By contrast, the Gini index is not decomposable in this way: it might happen that the inequality within a subgroup increases, but the Gini of the whole population diminishes, or vice versa. As a consequence, the Gini index is not appropriate for studying subgroups.[10]

Decomposability is a desirable property when it comes to use the indicators to analyse inequality within a population and its subgroups, and explore relations with other relevant variables, such as education or quality of life. However, lacking this property is not an error or a source of uncertainty because an indicator can have epistemic value independently of whether it is decomposable or not.[11]

Pragmatic considerations can influence several other choices. For example, when choosing the resource to focus on, researchers consider the feasibility of gathering data about it (McGregor et al., 2019). While income data is regularly collected by national statistical institutes and tax authorities, consumption data is not broadly available and requires costly surveys to be collected. Similarly, while financial wealth is relatively easy to measure, a more comprehensive estimation of someone's wealth is difficult because it requires bringing together many different data sources including national accounts, micro-level data from household surveys, financial institution surveys, administrative records, tax records, and specialized databases on artworks, luxury cars, etc. (McGregor et al., 2019).

3.4. Metrological

One last type of consideration is metrological. This driver appeals to the values of quantification and comparability. Quantification involves representing an attribute in terms specified by a mathematical scale, typically ratio, or interval.[12] Exactly how to formulate the conditions of each scale, what operations each allows, and how to estimate measurement uncertainty is the purview of metrology. Within the social sciences, metrological research is typically carried out as psychometrics, which uses mathematical models to represent the relation between psychological constructs and behavioural manifestations. Such models represent constructs as variables on particular scales (Borsboom and Molenaar, 2015). Quantification is valuable because it allows a consistent representation of attributes across various instances and enables the application of mathematical tools for making comparisons, as well as for deriving and testing predictions.[13] However, it is also controversial and difficult to defend especially in the human sciences.

For example, some have questioned whether parameters in the human sciences are genuinely quantifiable, either because they only justify ordinal rather than quantitative structures, or because they involve unjustifiable aggregations of heterogeneous properties (Michell, 2012; Hausman, 2015). Quantification in the human sciences is also criticized for oversimplifying social and psychological complexity, leading to a loss of interpretative depth and local contextual knowledge (Greene, 2020). These controversies have inspired work that examines the foundation of quantification in the human sciences and explores how measurement can be understood to account for its successful application in these fields (Basso, 2017; Briggs, 2021; Larroulet Philippi, 2021; Mari et al., 2023; Newfield, Alexandrova, and John, 2022; Vessonen, 2021a).

In the measurement of economic inequality, researchers usually disregard the question of whether representing inequality in quantitative terms is justified, presumably because economic inequality is predominantly measured in terms of income or wealth, which are quantified in monetary terms. However, the comparability of inequality across contexts and time periods is a central issue in methodological debates (Piketty, 2015; Sen, 1997a). This is because monetary values change over time and across currencies, and because the coexistence of similar but not identical indicators raises doubts about the comparability of their outcomes.

In particular, since inequality is estimated on the basis of the distribution of resources across a population, it is essential that economic circumstances are quantified in a way that allows for comparisons across the population. For example, when measuring inequality among households, researchers use equivalizing scales to improve the comparability of households of different numerosity and composition. Equivalizing scales are vectors of coefficients that standardise heterogeneous households and make them comparable. Multiple scales have been proposed in the literature and researchers are encouraged to choose the one that ensures the best comparability of available data and resources. Similarly, when measuring inequality among all individuals globally, regardless of residence, a key challenge is how to compare incomes in different currencies. There are alternative methods to do this, such as using Purchasing Power Parity (PPP) or current exchange rates to standardise monetary values. While none of these options is clearly better, researchers are advised to select the one that facilitates broader comparisons, other things being equal, and to cross-check their results using different methods.

Metrological considerations extend to other choices. In the absence of standardised procedures, each country measure inequality using its own methods and presuppositions, raising the problem of how to compare the outcomes of similar but not identical indicators (Atkinson and Brandolini, 2001, 2009). There is a rich methodological discussion about how to harmonize the outcomes of different indicators to make them comparable, without compromising their epistemic value (Jenkins, 2015; Basso 2025). Harmonization involves choosing a benchmark definition of inequality and constructing an indicator that is operable across a broad range of countries and desired timeframes.

Our goal in this section has been to emphasize that each driver creates opportunities for researchers to make different choices without determining them. With the four drivers in place, we are ready to introduce our case study in which the choices have already been made and need to be justified.

[10] Depending on the context of application, ignoring subgroups can have ethical implications, making this choice one that requires both ethical and pragmatic considerations.

[11] Note that it is possible to ask different questions about decomposability, such as if one is justified in representing inequality with a decomposable indicator, similarly for example to when researchers ask if temperature can be measured on an absolute scale. But economists do not discuss this question. They assume a positive answer and discuss when it is preferable to use a decomposable indicator, based on its pragmatic value for the purpose at hand.

[12] The metrological driver encompasses some of the considerations described by Bradburn et al. (2017) as part of the representational stage of measurement – when a researcher develops a metrical system to appropriately represent the property at hand. However, our proposal is not intended to describe stages of validation. As we will demonstrate in the following sections, all drivers contribute to the development and evaluation of measurement, and none takes precedence over the others, either temporally or philosophically.

[13] Mari et al. (2023, p. 259) characterize subject-independence (or intersubjectivity) in a way that partially captures these considerations, as it emphasises the goal that the information conveyed by a measurement should be interpretable in the same way by anyone.

4. The world inequality report 2022

The World Inequality Report 2022 is considered one of the best sources of information about global economic inequality. It is well-received in academic circles and broadly discussed in the media. The Report is authored by some of the most well-known economists of inequality in collaboration with a large network of researchers and institutions. It includes summaries and comments on recent and historical findings about global economic inequality, based on the database curated by the World Inequality Lab – an international institute based at the Paris School of Economics and the University of California, Berkeley. The Report describes today's levels of inequality as well as historical trends going back to 1820, covering almost all countries in the world.

The Report is the source of some striking claims. For example, it claims that "the poorest half of the world population owns just 2 % of total net wealth, whereas the richest half owns 98 % of the wealth on earth" (Chancel et al., 2022, p. 27). Another striking claim regards the disproportionate share of recent growth that is captured by global multimillionaires, contributing to an escalating trend of global inequality: "the top 1 % took 38 % of all additional wealth accumulated since the mid-1990s, whereas the bottom 50 % captured just 2 % of it" (ibid, p. 15).

The Report's findings not only expose unprecedented levels of inequality but also challenge widespread views about the causes of inequality. For instance, they challenge the ideas that inequality can be explained by geographical location or by average income levels, and that redistribution have large impacts on global inequality. By studying the impact of redistributive taxation, the Report states that while inequality is generally lower after taxation, the overall effect is small, so that in regions that are highly unequal before taxes, inequality remains extremely high even after redistribution. Moreover, the Report highlights that the redistributive power of taxation emerges as a historical contingency of the 19th and 20th centuries (when progressive tax systems have been established) and should not be taken for granted. Indeed, tax systems in several countries have become regressive in recent decades, meaning that the rich pay less tax, as a share of their income, than the middle and poorest part of the population (ibid, p.35).

To arrive at these finding, the Report investigates economic inequality using a variety of indicators, such as inequality pre- and post-taxation, inequality between and within countries, and across regions or globally regardless of where people live. These amount to multiple measurements of closely related parameters, which are used for causal analysis and hypothesis testing. For example, the Report measures inequality pre- and post-taxation to assess the impact of redistribution policies. Similarly, it looks at inequality within and between countries to track different dynamics of global inequality (Basso & Lisciandra, 2024). Despite the variety of measurements, some key methodological choices remain consistent throughout the Report. In particular, we discuss four key choices.

First, the Report focusses on inequality of income and wealth only, disregarding other factors that contribute to people's economic circumstances, such as public goods, human capital, or health.[14] Inequality of opportunities is also neglected. Second, the Report chooses to measure inequality using percentiles rather than summarizing statistics – that is, it focusses on which share of income and wealth is held by different parts of the population, as opposed to indexes summarizing the concentration of resources. Third, the Report chooses benchmark concepts of income and wealth inequality, which are used across all countries regardless of contextual differences. Finally, the Report is open access and has a transparent methodology, providing access to data-series as well as meta-data about methodological and analytical techniques.

All these choices are justified in methodological discussions that can be interpreted as appealing to the four drivers discussed in the previous section.

The epistemic driver motivates the provision of an empirical grasp on economic inequality, a social phenomenon that cannot be sized with eyes and ears. The WIR aims to shape our perception of global economic inequality by producing data series and elaborating them with statistical models and indicators. The authors set as their primary objective to document new findings about global inequality and its evolution. More precisely, they aim to bring out factual information that may be limited in scope but can be trusted because the same results could be gained by different researchers asking the same question. The authors acknowledge that some disagreement about inequality is unavoidable, particularly when the investigation reflects differing views on what the desirable level of inequality is, and the social policies and institutions required to achieve and sustain it. Nonetheless, they believe that "it is possible to agree about certain facts about inequality," implying an aspiration to keep disagreement over values separate from knowledge claims about inequality levels and trends (ibid, p. 22).[15]

The ethical driver pushes the authors to publicise socially relevant information concerning inequality, particularly by exposing the underserved privilege stemming from the extreme concentration of wealth in the hands of a small elite. According to the authors, informing about economic inequality is instrumental in enabling citizens to exercise their right to shape public decisions. Rejecting the idea that knowledge about inequality should remain confined to the realm of experts, the authors set as their "chief objective to contribute to the power of many" (ibid, p. 23). To this end, they strive to provide accessible information that can reach broad audiences, while maintaining transparency regarding data sources and methodologies to facilitate independent judgment. In essence, motivated by concern, indeed, anguish over today's inequality levels, the authors aim to equip their readers with the information they need to contribute to democratic debates, with the ultimate goal of catalysing actionable responses.

The pragmatic driver is at work to construct indicators with the data sources that are currently available. The Report is based on a database called World Inequality Database (WID), which compiles inequality records for several countries and years. The curators of the WID do not collect new empirical data themselves. Instead, the WID is constructed by elaborating on data collected by other institutes, including national statistical institutes and tax authorities. The WID combines multiple data sources to overcome some of the limitations of using one type of data only, but it remains constrained by the availability of data due to privacy, secrecy, tax evasion.

Finally, the metrological driver makes inequality comparable globally and over a long time span. The Report's ambition is to provide the most comprehensive findings on the current levels and historical evolution of income and wealth inequality, aiming to cover "almost all countries in the world over long time periods" (ibid, p. 22). According to the authors, this allows them "to present systematic data on inequality at the global level and to analyse how it has evolved over time" (ibid, p. 22).

These examples demonstrate that each of the four drivers are in operation in the Report. The authors of WIR deploy all four types of considerations in choosing how to measure and to represent economic inequality. But so far it remains possible that these drivers supply requirements that can be satisfied in a particular order, say with the epistemic driver being the dominant one. Equally, it remains possible

[14] In addition to income and wealth, the Report studies gender and environmental inequalities. However, these are studied as separate types of social disparities, rather than as contributing to people's economic circumstances.

[15] An emerging consensus in the values in science literature discussed in Section 3.2 holds that a neat separation between factual and value-laden claims is often impossible and undesirable. Some authors, however, maintain that researchers should still aim to eliminate value influences on scientific reasoning, even if this aim remains unattainable (Menon & Jacob, 2023).

that all four drivers can be satisfied at once, thus acting as necessary conditions for the adequacy of indicators. Now we are in the position to show that this is not so. The drivers bump into each other and force trade-offs.

5. Tensions and compromises

Sometimes drivers pull in opposite directions. For example, the decision to focus on the distribution of income and wealth may be justified pragmatically on the basis of data availability, but it has epistemic limitations: income and wealth fluctuate for many reasons, making them unreliable representations of people's economic circumstances. Similarly, measuring global inequality may be an ethical priority, but it poses metrological challenges, as international differences in currencies and economies make comparisons and hence quantification deeply uncertain. As a result, in some cases a driver can be satisfied only at the cost of another.

These trade-offs are ubiquitous in the WIR, but three such tensions are especially visible as the next subsections show. Moreover, when the authors of WIR need to balance different desiderata, there is no easy way to bring different aims into accordance, and no single correct answer as to which balance is best. In the absence of a standard, principled way of addressing the tensions, researchers make case-by-case decisions that weigh up pro and con, compromising between different drivers.

5.1. Epistemic vs. ethical driver

The first tension is between the epistemic and the ethical drivers. These drivers pull in opposite directions regarding the role of moral values in guiding researchers' choices. The epistemic driver calls for truthful, fact-based information about inequality. To meet the epistemic demands, the Report strives to be an apolitical resource in a pluralistic society, providing information that is free from bias and ideology. Conversely, the ethical driver urges an ethical stance to highlight the most troubling aspects of current inequality. The authors are committed to exposing undeserved privilege, aiming to empower their readers to participate in democratic decisions.

Conflicts between the epistemic and ethical aims of measurement have sometimes been acknowledged by historians and philosophers of science, who highlighted how measurements are socially and institutionally constructed, embedding social and political intentions that may compromise their objectivity and impartiality (Desrosières, 1998; Hacking, 1995; Kula, 1986; Porter, 1995). This tension is also evident in the context of scientific modelling. For example, Elliott & McKaughan (2014) claim that non-epistemic considerations can sometime take precedence over epistemic values in model evaluation, as models are assessed not only for their fit with the world but also for their suitability to meet the needs of their users (see also Diekmann and Peterson, 2013). In the case of the WIR, this tension manifests in the author's dual commitment to achieving factual accuracy while also taking responsibility for the societal impact of their research.

To balance these competing demands, the WIR sometimes prioritizes the epistemic driver and sometimes the ethical one. The goal is not to optimize both drivers simultaneously, but rather to allow both to contribute partially, striking a balance that is justifiable from their point of view. To illustrate, consider the following example. To meet the epistemic demands, the WIR aims to restrict its findings to claims that are robust against value disagreement. In contrast, to address the ethical demands, it represents these findings in a way that highlights the Report's ethical and political implications.

Because different statistical tools embed distinct value assumptions, the authors test the robustness of their main findings against multiple indicators. For example, when reconstructing the historical trends in income inequality, the Report tests the main findings using percentiles, T10/B50 ratios, and other indicators, including the Gini coefficient. The claim that todays' global inequality mirrors levels observed around 1900 and is substantially higher than in 1820 holds independently of the statistics employed (ibid, p.56).

Similarly, when analysing global wealth distribution, the Report uses Purchasing Power Parity (PPP) to compare economic circumstances across regions with different living costs. PPP rates are based on combining and weighting price indices in different areas of consumption to reflect people's purchasing needs across the world. This requires making value assumptions about how to aggregate and weight spending needs of different segments of the population (Reiss, 2008, Ch, 2). To ensure robustness, the WIR cross-checks the findings with alternative methods and argues that the extreme concentration of global wealth does not depend on the specific method used to compare resources across countries. When results diverge, it adopts the more cautious estimate, explicitly highlighting any discrepancies: "when measuring global wealth inequality using market exchange rates, rather than purchasing power parities, then there is ever more inequality: the bottom 50 % owns less than 1 % of total wealth and the global top 10 % nearly 82 % of it." (Chancel et al., 2022, p. 27).

When it comes to conveying their main findings, however, the authors no longer strive to be apolitical. Instead, they choose to represent their findings in a way that emphasises the extreme concentration at the top of the distribution. Rather than using summarizing statistics like the Gini or the Theil indexes, the authors prefer to Report their findings using percentiles or rations. Percentiles allows to monitor the shares of income and wealth of different parts of the distribution as in the claim that the bottom 50 % of the population owns only 2 % of global wealth. The Report uses indicators based on ratios like the Top 10/Bottom 50 (T10/B50), which measure how many times more the rich earn compared to the poorest half (ibid, p.31). Percentiles and ratios are more intuitive to understand, making it easier to reach broad audiences. Moreover, they highlight the tails of the distribution, helping to reveal the increasing concentration of resources at the top tail, which instead remains unclear looking at summarizing indexes.

It is worth clarifying that the use of percentiles is not an extra step outside the measurement process. The choice of indicator is a fundamental part measurement, as it creates a unique ordering of distributions based on their inequality. In the WIR, this choice is driven by ethical concerns including consideration of the consequences of how the results are received.

This choice diverges from disciplinary standards and partly contravenes the epistemic driver. The percentile approach makes values highly visible in the WIR, highlighting the authors' clear preference for more progressive taxation. Moreover, the use of percentiles imposes an epistemic constraint on the Report's users by emphasizing certain aspects of the distribution over others: it draws attention to top tail dynamics while providing little information about the rest of the distribution. Note that values are difference-makers in this case. Alternative ethical considerations, such as those motivated by the aim of informing and justifying national policy, would require representing the entire distribution, particularly the middle, which plays a primary role in social spending and redistribution. As a consequence, the epistemic and ethical demands are not jointly met. The Report addresses both only partially, striking a balance that can only be justified from their specific point of view.

5.2. Metrological vs. epistemic driver

The second tension involves the epistemic and the metrological drivers. These drivers generate opposing demands with respect to the contextual accuracy of inequality claims. The epistemic driver demands to provide information that is faithful to the real world, prioritizing descriptive adequacy throughout the WIR, also in the case of context dependent aspects of inequality. The metrological driver, instead, calls for global comparability, and this requires simplifying and flattening out contextual differences.

As with the previous case, these demands cannot be fully met

together. The trade-off between contextual accuracy and breadth of comparability is well known in the social sciences. Bradburn et al. (2017) argue that researchers are pulled in opposite directions when constructing measurements. Optimizing contextual accuracy often leads to a proliferation of context-specific measures. This makes accumulating knowledge difficult, pushing researchers to rely on concepts that may not perfectly fit specific contexts but allow for easier comparison. In the context of inequality, this tension is acknowledged by economists who argue that it's easier to compare inequality within subsets of relatively homogeneous countries, where data quality and measurement methods are similar. Comparability becomes more challenging as the groups become more heterogeneous (Jenkins, 2015).

While it may not be possible to fully satisfy both demands, the WIR strikes a balance, sometimes prioritizing comparability and sometimes emphasizing contextual accuracy. Driven by metrological considerations, the Report standardises inequality measurements across countries and over time. The benchmark concept of income is post-replacement, pre-tax income, which reflects gross income before taxation and after pension and retirement schemas. Wealth is defined as the sum of financial and non-financial assets net of debts. The statistical unit is adult individuals, with resources equally split between couples. According to the authors, it is essential to use single benchmark concepts to avoid misleading conclusions in international comparisons.

This standardisation inevitably introduces distortions, partly compromising the epistemic ambitions. By using the same concepts across countries and over time, the Report flattens out contextual differences, sacrificing some descriptive adequacy for the sake of comparability and knowledge accumulation. For example, splitting household income equally between couples may not accurately reflect people's economic circumstances depending on legislation or demographic microstructures. In regions and time periods where women are not allowed to own land, splitting household resources equally bring about a distorted picture of the real distribution (Basso 2025).

The authors of WIR are aware of these dangers and seek to reduce the degree of distortion. To redeem descriptive adequacy at least to some extent, the Report takes into account some contextual differences. For example, when estimating national incomes, the Report considers country-specific definitions of taxable income, including the treatment of special forms of income like dividends, pensions, and social benefits. Similarly, when comparing average income levels between countries, the Report takes into account income earned per hour, as working hours vary significantly across countries. As a result, the claim that "North Americans earn 6 to 10 times more, on average, than Sub-Saharan Africans, South and Southeast Asians" is made more precise by adding that Sub-Saharan Africans and Southeast Asians spend around 30 % more time at work per year than Europeans and North Americans (Chancel et al., 2022, p. 28). By considering these contextual features, the WIR sacrifices some comparability but gains the advantage of correcting some relevant distortions. Once again, the Report cannot optimize both desiderata at the same time but strikes a balance between these competing needs.

5.3. Epistemic and ethical vs. pragmatic and metrological drivers

The third tension is more complex as it pits a pair of drivers against another pair. The epistemic and ethical drivers together reflect a commitment to make inequality visible and actionable. The Report aims to grasp empirically what is otherwise a nebulous social phenomenon, making it knowable and subject to intervention. Call this commitment *visibility*. Conversely, pragmatic and metrological drivers require the adoption of simple, technical definitions that are quantifiable and operable across countries and over time. The commitment of quantification is *fruitfulness* because it enables comparison and precise testing, but it also requires simplifying and narrowing down the notion of inequality. Neglected factors become invisible in the measurement outcomes, creating the need to balance visibility with fruitfulness of the indicators.

Historians of science and political philosophers have explored this tension by reconstructing how indicators evolved to fulfil political functions, such as capturing and communicating socially significant changes (Bach and Morgan, 2020) or serving as tools for justifying political decisions (Badano, 2022). These scholars emphasize that measurement has a positive function in addressing social problems and supporting public justification. However, the required simplifications are not without risks because they can obscure important factors, leaving measurements vulnerable to manipulation and misrepresentation.

To see how the WIR navigates this tension, consider the choice of focussing on income and wealth only. The focus on income and wealth facilitates quantification and comparability because it allows to draw on some of the best sources of data about economic inequality. However, these indicators offer only a limited perspective, ignoring other factors that influence people's economic circumstances. Education, public goods, and inequality of opportunity, for instance, are excluded, compromising visibility for the sake of quantification. Amartya Sen famously criticised the narrow understanding of economic inequality in terms of income only. He argued that income is merely one means among others for achieving personal and social goals, illustrating this with examples of countries where lower average incomes coexist with longer life expectancy thanks to factors like environmental conditions, social climate, and intrahousehold distribution (Sen, 1997b).

To make up for these limitations and broaden the scope of the analysis, the Report complements the study of income and wealth with additional indicators, such as time spent at work, quality of public services and infrastructure, and quality of civic and human rights (Chancel et al., 2022, p. 28). According to the authors, "incomes are a powerful economic indicator of living standards, but must be complemented by other indicators […] if they are to be a good representation of inequalities in living standards between countries" (ibid, p.28). The Report also provides extensive empirical analysis of other social inequalities, looking at income inequalities between genders and global ecological inequality. However, it offers little examination of how different social inequalities are related to each other.

Neither visibility, nor fruitfulness is fully achieved. In some cases, fruitfulness takes precedence, while in others, visibility is prioritised. The choice to focus on income and wealth as economic resources addresses the need for a fruitful quantification that can be used for comparison and precise testing, but this narrow focus limits visibility. On the other hand, using complementary indicators enriches the visibility and accountability of various aspects of inequality but complicates quantification. Both demands are only partially met, with the Report striking a balance between these incompatible needs.

6. Conclusions

We have argued that to understand measurements of inequality in the WIR is to understand the compromises between the epistemic, ethical, pragmatic, and metrological drivers. This Report is an example of social measurement that lends itself particularly well to this sort of analysis, but we trust that the idea of the four drivers and compromises between them is generalisable and applicable to other cases in the human and possibly also life sciences.

When constructing indicators, researchers rely on multiple types of considerations, which we loosely classify into four different drivers. These drivers can sometime be in tension to each other, creating competing demands.

Our examination suggests that there is no single correct way to address these tensions and no pre-established norms on how to resolve them. In the WIR, these tensions are not always resolved in favour of one driver over another. There is no temporal or priority order among the drivers, nor are there clearly distinct stages of research where different drivers apply. Instead, researchers' choices should be understood as a series of compromises, tailored to the specific circumstances.

We also find that none of the drivers plays a consistently secondary role. Ethical and pragmatic considerations, which are usually regarded as potential disturbances or treated as secondary factors that only come into play when choices are epistemically underdetermined, hold a more substantial role in WIR. These considerations not only ensure ethical admissibility and practical feasibility. They also shape decisions about balancing epistemic accuracy and quantification, but in a way that is constrained by the ambition to state the facts. No single ambition has lexical priority over the others.

Recognising this balancing act matters for how we assess measurement quality. First, it challenges the idea that this quality can be evaluated along a single dimension, a message concordant with the contextual conception of validity in Larroulet Philippi (2021). When measurement involves balancing multiple, often conflicting demands, constructing a good indicator is not just about optimizing its epistemic performance. Instead, it requires finding a defensible balance between different needs and demands. The quality of an indicator, its ability to serve its intended purposes, and its authority for decision making processes ultimately depend on compromises between the drivers. We thus urge a realistic view of measurement, which acknowledges the inevitability of sacrifices and encourages researchers to be explicit about their choices.

A second implication emerges through an analogy with the evaluation of scientific models. Parker (2020) argues that models should be assessed based on their adequacy for a specific purpose, rather than solely on how accurately and completely they represent a target system. This perspective emphasises that models, like measurements, are not just representations but also tools designed to fulfil both epistemic and practical purposes, with their usefulness and persuasiveness varying depending on the needs and perspectives of users (Giere, 2004; Knuuttila, 2011; Mäki, 2011; see Rolin, 2024 for an overview). In our case, adequacy for purpose is a high-level representation of the compromises between different drivers.[16]

The compromises that are deemed acceptable when evaluating an indicator depend on the purposes the indicator is meant to serve. For example, we noted that the Report's emphasis on the tails of the distribution is motivated by the ethical driver, at the expense of fully capturing trends and dynamics across the entire distribution. This compromise appears to align with the purposes of the WIR, which aims to highlight global distribution dynamics for broad audiences and potentially guide international cooperation and foreign aid. However, such a compromise might not be suitable for other purposes. For instance, informing and justifying national policy would likely require a focus on the entire distribution, especially the middle, which is central to decisions about social spending and redistribution. We hypothesize that if the Report were designed to inform national policymaking, the compromise prioritizing the tails of the distribution would likely seem less acceptable.

As with many aspects of science, measurement involves compromises. No compromise is valid universally, but some are much more defensible than others.

CRediT authorship contribution statement

Alessandra Basso: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Anna Alexandrova:** Writing – review & editing, Supervision, Conceptualization.

[16] Parker and Winsberg (2018, p.129) noted that, since new models are often built on earlier ones, the purposes of past modellers can continue to exert influence. This perspective can be relevant for inequality measurement too. For example, the Gini index, first developed by Corrado Gini in 1912, was based on Max Lorenz's earlier model of income distribution, the Lorenz Curve. Over time, further refinements and alternative formulations have been proposed, such as those by Jasso (1979) and Deaton (1997).

Declaration of interest

The authors declare no conflict of interest.

Acknowledgments

The authors gratefully acknowledge the valuable comments received from participants at the 2023 conference of the International Network for Economic Method (INEM) and the workshop on *Measuring the Human* (Cambridge, July 2023). We are also indebted to two anonymous referees for their thoughtful and constructive suggestions. This research was supported by funding from The British Academy and The Isaac Newton Trust.

References

Afonso, H., LaFleur, M., & Alarcón, D. (2015). Inequality measurement. *Develoment Issues No. 2 United Nations' Development Strategy and Policy Analysis Unit in the Development Policy and Analysis Dividion of UN/DESA*.

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

Alexandrova, A., & Fabian, M. (2022). Democratising measurement: Or why thick concepts call for coproduction. *European Journal for Philosophy of Science, 12*(1), 7. https://doi.org/10.1007/s13194-021-00437-7

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia, 19*(1), 1–24. https://doi.org/10.2979/hyp.2004.19.1.1

Angner, E. (2013). Is it possible to measure happiness? *European Journal for Philosophy of Science, 3*(2), 221–240. https://doi.org/10.1007/s13194-013-0065-2

Atkinson, A. B. (1983). *The economics of inequality*. Oxford: Clarendon Press.

Atkinson, A. B., & Brandolini, A. (2001). Promise and pitfalls in the use of "secondary" data-sets: Income inequality in OECD countries as a case study. *Journal of Economic Literature, 39*(3), 771–799.

Atkinson, A. B., & Brandolini, A. (2009). On data: A case study of the evolution of income inequality across time and across countries. *Cambridge Journal of Economics, 33*(3), 381–404. https://doi.org/10.1093/cje/bel013

Auten, G., & Splinter, D. (2023). Income inequality in the United States: Using tax data to measure long-term trends. *Journal of Political Economy*, (November), Article 728741. https://doi.org/10.1086/728741

Bach, M., & Morgan, M. S. (2020). Measuring difference? The united nations' shift from progress to poverty. *History of Political Economy, 52*(3), 539–560. https://doi.org/10.1215/00182702-8304855

Badano, G. (2022). Are numbers really as bad as they seem? A political philosophy perspective. In C. Newfield, A. Alexandrova, & S. John (Eds.), *Limits of the Numerical: The Abuses and Uses of quantification* (pp. 161–178). University of Chicago Press. https://doi.org/10.7208/chicago/9780226817163-008.

Baldini, M., & Toso, S. (2009). *Diseguaglianza, povertà e politiche pubbliche* (New ed). Bologna: Il mulino: Itinerari Economia.

Basso, A. (2025). The Comparison of Inequality Measurements across Countries and Time. *The British Journal for the Philosophy of Science, 76(3)*. https://doi.org/10.1086/719568

Basso, A. (2017). The appeal to robustness in measurement practice. *Studies in History and Philosophy of Science Part A, The Making of Measurement, 65–66*(October), 57–66. https://doi.org/10.1016/j.shpsa.2017.02.001

Basso, A., & Lisciandra, C. (2024). Models and measurement of inequality. In T. Knuuttila, N. Carrillo, & R. Koskinen (Eds.), *Routledge Handbook of scientific modeling* (pp. 511–524). Routledge.

Bojer, H. (2003). *Distributional justice: Theory and measurement* (Vol. 47). London ; New York: Routledge Frontiers of Political Economy. Routledge.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. D. Wright (Ed.), *International Encyclopedia of the social & Behavioral sciences* (2nd ed., Vol. 19, pp. 418–422). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.43079-5.

Boumans, M. (2015). *Science outside the Laboratory: Measurement in field science and economics*. Oxford, New York: Oxford University Press.

Bradburn, N., Cartwright, & Fuller, J. (2017). A theory of measurement. In L. McClimans (Ed.), *Measurement in Medicine : Philosophical Essays on assessment and evaluation*. London: Rowman & Littlefield.

Briggs, D. C. (2021). *Historical and Conceptual foundations of measurement in the human sciences: Credos and controversies* (1st ed.). New York, NY: Routledge.

Bruno, I., Didier, E., & Vitale, T. (2014). Statactivism: Forms of action between disclosure and affirmation. *The Open Journal of Sociopolitical Studies, 7*(2), 198–220.

Cartwright, & Runhardt, R. (2014). Measurement. In N. Cartwright, & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (pp. 265–287). Oxford: Oxford University Press.

Case, A., & Deaton, A. (2020). *Deaths of despair and the Future of capitalism*. Princeton: Princeton University Press.

Chancel, L., Piketty, T., Saez, E., Zucman, G., Bajard, F., Burq, F., Moshrif, R., Neef, T., & Robilliard, A.-S. (2022). World inequality report 2022. *World Inequality Lab. wir2022.wid.world*.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Chang, H., & Cartwright, N. (2008). Measurement. In S. Psillos, & M. Curd (Eds.), *The Routledge Companion to philosophy of science* (pp. 367–375). New York: Routledge.

Costa, R. N., & Pérez-Duarte, S. (2019). Not all inequality measures were created equal: The measurement of wealth inequality, its Decompositions, and an application to European household wealth. *ECB Statistics Paper*. https://doi.org/10.2866/957288

Coulter, F. A. E., Cowell, F. A., & Jenkins, S. P. (1992). Equivalence scale relativities and the extent of inequality and poverty. *The Economic Journal, 102*(414). https://doi.org/10.2307/2234376

Deaton, A. (1997). *The analysis of household surveys: A Microeconometric approach to development policy*. World Bank Publications.

Desrosières, A. (1998). *The Politics of large numbers: A History of statistical reasoning. Translated by Camille naish*. Cambridge, MA: Harvard University Press.

Diekmann, S., & Peterson, M. (2013). The role of non-epistemic values in engineering models. *Science and Engineering Ethics, 19*(1), 207–218. https://doi.org/10.1007/s11948-011-9300-4

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. Pittsburgh, Pa: University of Pittsburgh Press.

Drimmelen, T. van, Nienke Slagboom, M., Reis, R., Bouter, L. M., & van der Steen, J. T. (2024). Decisions, decisions, decisions: An ethnographic study of researcher discretion in practice. *Science and Engineering Ethics, 30*(6), 59. https://doi.org/10.1007/s11948-024-00481-5

Elliott, K. C. (2011). *Is a little Pollution good for You? Incorporating societal values in environmental research. Environmental ethics and science policy series*. New York: Oxford university press.

Elliott, K. C., & McKaughan, J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science, 81*(1), 1–21.

Fraassen, B. C. van (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*(5), 742–752. https://doi.org/10.1086/425063

Giordani, A., & Mari, L. (2013). Modeling measurement: Error and uncertainty. In M. Boumans, G. Hon, & A. Petersen (Eds.), *Error and uncertainty in scientific practice* (pp. 79–96). Pickering & Chatto.

Greene, C. (2020). Nomadic concepts, variable choice, and the social sciences. *Philosophy of the Social Sciences, 50*(1), 3–22. https://doi.org/10.1177/0048393119878783

Hacking, I. (1995). The Looping effects of human kinds. In C. Cognition, & A. Multidisciplinary Debate (Eds.), *Symposia of the fyssen foundation* (pp. 351–394). New York, NY, US: Clarendon Press/Oxford University Press.

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. Wiley.

Hausman, D. M. (2015). *Valuing health: Well-being, freedom, and Suffering. Population-level Bioethics series*. Oxford ; New York: Oxford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jasso, G. (1979). On Gini's mean difference and Gini's index of concentration. *American Sociological Review, 44*(5), 867–870. https://doi.org/10.2307/2094535

JCGM, (Joint Committee for Guides in Metrology). (2020). *Guide to the expression of uncertainty in measurement*.

Jenkins, S. (2015). World income inequality databases: An assessment of WIID and SWIID. *The Journal of Economic Inequality, 13*(4), 629–671. https://doi.org/10.1007/s10888-015-9305-3

Jenkins, S., Van Kerm, P., & Salverda, W. (2009). The measurement of economic inequality. In B. Nolan, M. Timothy, & Smeeding (Eds.), *The measurement of economic inequality*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199606061.013.0003.

Jenkins, S. P. (2024). Getting the measure of inequality. *Oxford Open Economics, 3*(1), 156–166. https://doi.org/10.1093/ooec/odad037

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science Part A, Model-Based Representation in Scientific Practice, 42*(2), 262–271. https://doi.org/10.1016/j.shpsa.2010.11.034

Foundations of measurement. In Krantz, D., Luce, D., Suppes, P., & Tversky, A. (Eds.), *Additive and Polynomial representations*, (1971). New York: Academic Press.

Kuhn, T. S. (1981). Objectivity, value judgment, and theory choice. In D. Zaret (Ed.), *Review of Thomas S. Kuhn the essential tension: Selected studies in scientific tradition and change* (pp. 320–339). Duke University Press.

Kula, W. (1986). *Measures and men*. Princeton, N.J: Princeton University Press.

Lakner, C., & Milanovic, B. (2016). Global income distribution: From the fall of the Berlin wall to the great Recession. *The World Bank Economic Review, 30*(2), 203–232.

Larroulet Philippi, C. (2021). Valid for what? On the very idea of unconditional validity. *Philosophy of the Social Sciences, 51*(2), 151–175. https://doi.org/10.1177/0048393120971169

Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–431.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications: Revisions, new directions and applications* (Illustrated edition). Charlotte, NC: Information Age Publishing.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton, N.J: Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton, N.J: Princeton University Press.

Lusk, G., & Elliott, K. C. (2022). Non-epistemic values and scientific assessment: An adequacy-for-purpose view. *European Journal for Philosophy of Science, 12*(2), 35. https://doi.org/10.1007/s13194-022-00458-w

Mäki, U. (2011). Models and the locus of their truth. *Synthese, 180*(1), 47–63.

Mari, L., Wilson, M., & Maul, A. (2023). Measurement across the sciences: Developing a shared concept system for measurement. *Springer Nature*. https://doi.org/10.1007/978-3-031-22448-5

McClimans, L. (2010). A theoretical framework for patient-reported outcome measures. *Theoretical Medicine and Bioethics, 31*(3), 225–240. https://doi.org/10.1007/s11017-010-9142-0

McClimans, L. (2024). *Patient-centered measurement: Ethics, Epistemology, and Dialogue in Contemporary Medicine. Oxford scholarship online*. New York, NY: Oxford University Press. https://doi.org/10.1093/oso/9780197572078.001.0001

McGregor, T., Smith, B., & Wills, S. (2019). Measuring inequality. *Oxford Review of Economic Policy, 35*(3), 368–395. https://doi.org/10.1093/oxrep/grz015

Menon, T., & Jacob, S. (2023). Sisyphean science: Why value freedom is worth pursuing. *European Journal for Philosophy of Science, 13*(48), 1–24. https://doi.org/10.1007/s13194-023-00552-7

Michell, J. (2012). "The constantly recurring argument": Inferring quantity from order. *Theory & Psychology, 22*(3), 255–271. https://doi.org/10.1177/0959354311434656

Morgan, M. S., & Bach, M. (2018). Measuring development—from the UN's perspective. *History of Political Economy, 50*(S1), 193–210. https://doi.org/10.1215/00182702-7033932

Newfield, C., Alexandrova, A., & John, S. (Eds.). (2022). *Limits of the Numerical: The Abuses and Uses of quantification*. Chicago, IL: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/L/bo146791774.html.

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science, 87*(3), 457–477. https://doi.org/10.1086/708691

Parker, W. S., & Winsberg, E. (2018). Values and evidence: How models make a difference. *European Journal for Philosophy of Science, 8*(1), 125–142. https://doi.org/10.1007/s13194-017-0180-6

Peschard, I. F., & van Fraassen, B. C. (2014). Making the abstract concrete: The role of norms and values in experimental modeling. *Studies In History and Philosophy of Science Part A, 46*, 3–10. https://doi.org/10.1016/j.shpsa.2013.11.004

Piketty, T. (2015). *The economics of inequality*. Cambridge, Massachusetts ; London, England: The Belknap Press of Harvard University Press.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, New Jersey: Princeton University Press.

Potochnik, A. (2020). Idealization and many aims. *Philosophy of Science, 87*(5), 933–943. https://doi.org/10.1086/710622

Ray, D. (1998). *Development economics*. Princeton, N.J: Princeton University Press.

Reiss, J. (2008). Measurement between the absolute and the arbitrary. In *Error in economics*. Routledge.

Rodriguez Duque, S., Tal, E., & Barbic, S. P. (2024). The role of ethical and social values in psychosocial measurement. *Measurement, 225*(February), Article 113993. https://doi.org/10.1016/j.measurement.2023.113993

Rolin, K. (2024). Models and values. In T. Knuuttila, N. Carrillo, & R. Koskinen (Eds.), *The Routledge Handbook of philosophy of scientific modeling*. Routledge.

Rosales, A., & Morton, A. (2019). Scientific explanation and trade-offs between explanatory virtues. *Foundations of Science, 26*(4), 1075–1087. https://doi.org/10.1007/s10699-019-09645-0

Salverda, W., Nolan, B., & Smeeding, T. M. (Eds.). (2009). *The Oxford Handbook of economic inequality*. Oxford ; New York: Oxford University Press.

Schroeder, S. A. (2019). Which values should Be built into economic measures? *Economics and Philosophy, 35*(3), 521–536. https://doi.org/10.1017/s0266267118000317

Sen, A. (1997a). In J. E. Foster, & A. Sen (Eds.), *On Economic Inequality. Expanded edition with a substantial annexe*. Oxford : New York: Clarendon Press ; Oxford University Press.

Sen, A. (1997b). Distinguished Guest Lecture: From Income Inequality to Economic Inequality. *Southern Economic Journal, 64*(2), 384–401.

Simmons, J. P., Nelson, L. D., & Uri, S. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science, 77*(1), 14–34. https://doi.org/10.1086/650206

Tal, E. (2016). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science, 67*(1), 297–335. https://doi.org/10.1093/bjps/axu037

Tal, E. (2017). A model-based epistemology of measurement. In N. Mößner, & A. Nordmann (Eds.), *Reasoning in measurement*. London: Routledge.

Thoma, J. (2024). Social science, policy and democracy. *Philosophy & Public Affairs, 52*(1), 5–41. https://doi.org/10.1111/papa.12250

Vessonen, E. (2021a). Conceptual engineering and operationalism in psychology. *Synthese, 199*(3), 10615–10637. https://doi.org/10.1007/s11229-021-03261-x

Vessonen, E. (2021b). Representation in measurement. *European Journal for Philosophy of Science, 11*(3), 76. https://doi.org/10.1007/s13194-021-00365-6

Vet, H. C. W. de, Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in Medicine: A practical guide. Practical guides to Biostatistics and Epidemiology*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511996214

Zeller, R. A., & Carmines, E. G. (1980). *Measurement in the social sciences: The link between theory and data*. Cambridge ; New York: Cambridge University Press.