Faculty of Social Sciences
University of Helsinki

Measuring the social world

Studies in the epistemology of measurement in the social and behavioural sciences

Alessandra Basso

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Social Sciences of the University of Helsinki, remotely, on the 22nd of April, 2022 at 13 o' clock.

Helsinki 2022

ISBN 978-951-51-7049-1 (paperback)
ISBN 978-951-51-7050-7 (pdf)
ISSN 2343-273X (print series)
ISSN 2343-2748 (web series)
Unigrafia 209/2022
Helsinki 2022

Abstract

This dissertation investigates measurement practice in the social and behavioural sciences in order to address the challenges that arise when measuring phenomena in these fields.

The studies in this dissertation argue that there is continuity in the notion of measurement accuracy across the natural, social, and behavioural sciences because measurement assessment practice, and the notion of accuracy it relies upon, exhibits important similarities across disciplines. As with physical measurement, social and behavioural scientists assess the reliability of measurement by looking at the robustness of the outcomes across multiple measurements of the same parameter. Consistency across different measurements is taken as a sign that the outcomes can be attributed to what is measured. Scientists can improve this consistency by detecting and correcting for errors. In this assessment practice, the accuracy of measurement is the degree of agreement among multiple measurements of the same parameter.

However, due to the nature of the phenomena being studied, social and behavioural measurement involves a greater degree of approximation and requires more flexibility in the measurement procedures and assessment methods. Many of the phenomena studied in the social and behavioural sciences are multifaceted and/or loosely defined and therefore some measurement procedures cannot be modelled to the same level of detail as the measurement of well-defined physical quantities. As a result, the measurement of these phenomena necessarily involves greater degrees of approximation. Social and behavioural phenomena, moreover, often change over time and have different characteristics across contexts, and as such might require partially distinct kinds of measurement. Additionally, unlike in the physical sciences, social and behavioural phenomena are often morally laden and as a consequence their measurement requires making value judgments, which are highly sensitive to the context of application. Therefore it is not obvious that a standardised definition of these complex and changing phenomena is the best epistemic strategy to pursue. When measuring these phenomena, scientists face a trade-off between the need to standardise measurement practice in order to improve comparability and facilitate the accumulation of knowledge, and the demand to tailor the definition to the specific contexts being studied.

In my view, context-dependency and approximation do not lead to pessimistic conclusions about the reliability of measurement in these fields, but should rather be seen as guiding measurement practice and the employment of its outcomes. For instance, to deal with context-dependency, scientists can adopt strategies that trace regularities in context-dependence and improve the comparability across context-dependent measurements. When the measurement is influenced by the purpose and context of application, what counts as good measurement is also sensitive to purpose and context, and therefore can admit different degrees of approximation. The studies in this dissertation suggest ways to improve how scientists deal with these challenges and warn against ways of using the outcomes that can raise conceptual or epistemological doubts.

Acknowledgments

First and foremost, I would like to thank my supervisors Uskali Mäki, Caterina Marchionni and Eran Tal. Uskali's enthusiasm for philosophy and his deep thinking were contagious from the first time we met. His ability to zoom in and out of philosophical reflection, as well as uncover philosophically interesting questions in the most diverse contexts, is unparalleled. I appreciate his encouraging words and energetic support. I am also grateful to Uskali for bringing together a fantastic research community with whom I had the pleasure of collaborating during my time in Helsinki.

I am extremely grateful to Caterina Marchionni for her guidance and example, which played an essential role in my intellectual development since the beginning of my doctoral studies. Her keen observations have greatly contributed to my articles and to the overall argument of this dissertation. Her philosophical knowledge and her friendship are enormously valuable to me. I learned from co-authoring articles with her, as she tirelessly provided insightful comments and practical suggestions. She joined the supervising team near the end of my doctoral studies, when supervision was most needed. Her constant feedback and patient advice helped me in the final stages of my PhD.

I am most indebted to Eran Tal. I met Eran during my first year of doctoral studies and asked him to be my supervisor about halfway through. He had the greatest impact on the content of this dissertation. I have benefited greatly from Eran's intellectual honesty and many keen suggestions, which have helped me develop and refine my arguments. Without Eran's extensive knowledge of philosophical thought about measurement, this dissertation would have been significantly less than it is now. I could not hope for better support and consider myself privileged to have had such a dedicated and encouraging team of supervisors.

I would like to thank my thesis committee members, Petri Ylikoski and Tarja Knuuttila for helpful tips and guidance from a remote meeting that brought Padua, Helsinki and Vienna into the same virtual room.

A special thank you goes to Simo Kyllönen, co-author of one the dissertation articles, for enriching discussions and a genuine attempt to bring together different streams of philosophical thought.

During my doctoral studies, I had the honour of being part of a vibrant research group in Helsinki and the wider philosophical community that surrounds it. The TINT Centre for Philosophy of Social Science has been a stimulating research environment that provided constant inspiration, insightful feedback, and discussion. I would like to thank everyone who has been a part of this community over the years. Thank you in particular to Aki Lehtinen, Jaakko Kuorikoski, Samuli Reijula, Marion Godman, Säde Hormio, Luis Mireles Flores, Carlo Martini, Michiru Nagatsu, Chiara Lisciandra, Till Grüne-Yanoff, Magdalena Małecka and many others who have provided constructive feedback and discussion.

Because of the exceptionally difficult times we have been through, I am extraordinarily grateful to family and friends. Writing a dissertation amid a pandemic, with children home from school and socially isolated, requires a lot of help. I want to express my gratitude to my brother Michele and his wife Emma for offering a place to

study while everything else was closed, complete with coffee, food, and of course good company. My parents deserve a heartfelt thank you for loving their grandchildren during times of social isolation and providing them with all the care and affection they needed. A special thank you goes to my mother for her unwavering support of this project, even when I was short of inspiration.

Finally, I wish to thank my children, Giorgio and Elia, for the most enjoyable playful time and above all for being an endless source of love and happiness.

Contents

List of original publications

This dissertation consists of the following publications:

I. **The appeal to robustness in measurement practice**
Alessandra Basso
Studies in History and Philosophy of Science, 65(57-66), 2017.

II. **The comparison of inequality measurements across countries and time**
Alessandra Basso
Forthcoming *in The British Journal for the Philosophy of science.*

III. **From measurement to classificatory practice: improving psychiatric classification independently of the opposition between symptom-based and causal approaches**
Alessandra Basso
European Journal for Philosophy of Science, 11(104), 2021.

IV. **When Utility Maximisation is not Enough: Sufficientarianism and the Economics of Climate Change**
Simo Kyllönen and Alessandra Basso
In Walsh, A., Hormio, S. and Purves, D. (eds.) *Ethical Underpinnings of Climate Economics,* Routledge 2017.

The publications are referred to in the text by their roman numerals.

Part I: Introductory essay

1. Introduction. Measurement practice across disciplinary boundaries

Measurement is praised for providing standardised procedures that can be used across contexts and for producing results that can be interpreted in the same way by different subjects, thereby facilitating the comparison and accumulation of knowledge. With its application to the physical sciences, measurement came to be considered a reliable source of knowledge. The application of measurement in the social sciences, but also in psychology and psychiatry,[1] however, is much more contested and frequently provokes heated debates about what is being measured and how.

In the social and behavioural sciences, the definitions of the parameters being measured rarely enjoy a level of consensus comparable to the general acceptance of well-defined physical quantities. One reason is that many of the phenomena studied in these fields, like poverty or quality of life, are characterised by clusters of features with no clear boundaries, and this makes it difficult to come up with a precise definition of these phenomena. For example, quality of life has multiple dimensions, such as psychological, material, and environmental ones, and it is characterised by an open-ended list of features like absence of pain, freedom, social mobility, and wealth. Moreover, the phenomena studied in these disciplines often change over time and have different characteristics across contexts, and therefore might require partially distinct kinds of measurement. For instance, what counts as *being in poverty* can vary greatly across countries. Standardising the measurement of such complex and changing phenomena might not be the optimal epistemic strategy, because it would require foregoing many of its context-dependent characteristics. Standardisation, in fact, requires choosing a definition to be used across contexts, and therefore a standardised measurement is inadequate for capturing features that vary depending on the context of application. When measuring context-dependent phenomena, scientists face a trade-off between two competing desiderata: standardising the measurement in order to improve comparability and facilitate the accumulation of knowledge, and tailoring the measurement to the context under study to capture what is most relevant for the purpose at hand.

Another set of challenges regards the assessment of social and behavioural measurement. When the phenomena being measured are complex and context-dependent, scientists might be unable to meet the conditions for repeated observations, or to achieve the same kind of controlled variations as in the evaluation of well-defined physical measurement. As a result, the assessment of measurement in these disciplines involves a greater degree of approximation. Finally, social and behavioural scientists are often interested in measuring morally charged phenomena like poverty, inequality and well-being. The measurement of these phenomena requires making value judgments. For instance, measuring inequality requires making

[1] In what follows, I refer to this group of disciplines as the social and behavioural sciences. It includes the social sciences, various branches of psychology, and also psychiatry and large parts of medicine.

choices about what distribution of resources is desirable. These choices impact on the reliability of the measurement and its scope of application and, as a consequence, the measurement of morally charged phenomena can be criticised on ethical grounds. Measurement in the social and behavioural sciences thus raises specific epistemological questions, such as 'how is it possible to measure parameters that are context-dependent and morally charged?'; 'what is the best way to conceptualise these parameters for the purpose of measurement?'; "how can one tell whether the outcomes provide information on what is measured and how well do they do so?'; and 'how much of the social world can in fact be measured?'

This dissertation contributes to addressing these questions by studying measurement practice in the social and behavioural sciences. There are two main motivations for investigating this topic. First, measurement in these fields is largely neglected in the philosophical literature, which instead focuses mainly on physical measurement. This is a drawback for improving the quality and the social impact of measurement in these disciplines. Driven by the demands of evidence-informed policymaking and evidence-based medicine, the application of measurement in the social and behavioural sciences is increasingly expected to arbitrate in the design, selection, and implementation of policy in areas ranging from health to economic development. However, the reliability of these measurements is contested, and therefore their impact on science and policymaking has also been challenged. The investigation of the epistemological basis underlying measurement practice in these fields has great potential both to contribute to understanding measurement practice outside the physical realm, and to improve social and behavioural measurement and promote the meaningful employment of its outcomes.

Second, the focus on measurement practice provides a new way of looking at the epistemological justification of measurement in these fields and the conditions under which it is reliable. This approach, therefore, promises novel solutions to the challenges that arise when measuring social and behavioural phenomena. Indeed, it is only recently that philosophers have turned their attention to measurement practice and the various tasks it involves. The philosophical literature on measurement has traditionally focused on topics like the metaphysics of quantities, the semantics of measurement, and the mathematical foundations of scales. Issues related to the conditions that make measurement reliable and practical strategies to evaluate and improve its accuracy remained relatively unexplored. Contemporary literature on the 'epistemology of measurement' have instead begun to engage with measurement practice and the various tasks it involves, such as the definition of the parameters, instrument design and calibration, and especially the correction of errors and the evaluation of measurement accuracy (Chang 2004; van Fraassen 2008; Tal 2011; 2016; 2019; Frigerio, Giordani, and Mari 2010; Boumans 2006; 2015). Some of these philosophical works, moreover, have elaborated model-based accounts of measurement that emphasise the role of theoretical and statistical models in measurement.

The studies included in this dissertation address some of the challenges that arise when measuring social and behavioural phenomena by clarifying the conceptual and epistemological presuppositions underlying measurement practice in these fields. The account of measurement practice that emerges from these studies emphasises the role of representation in the development of measurement and the interpretation of

its outcomes. While measurement always involves a concrete procedure for assigning values to the parameter of interest, the outcomes also depend on the way in which the parameter and the measurement procedure are represented by models. Indeed, social and behavioural scientists gather empirical indications such as survey responses and interviews, but these indications need to be corrected and interpreted on the basis of a model of how the measurement works. For example, scientists collect and analyse survey data to obtain values for parameters like unemployment, living standards, and inequality, but this data must be revised on the basis of theoretical assumptions about the parameter of interest and the population under study. When there are reasons to assume that individual responses are influenced by systematic sources of error, scientists correct the outcomes for misreporting.

Rather than drawing pessimistic conclusions about the reliability of measurement in the social and behavioural sciences, this dissertation takes a constructive approach to deal with the challenges that arise in these disciplines. The dissertation argues that scientists can mitigate some of the problems of social and behavioural measurement by using specific methodologies and suggests ways to evaluate and improve their strategies. In particular, the dissertation articles find that the inferential pattern underlying measurement assessment, and the empirical conditions for reliability, shares some similarities between the natural, the social and the behavioural sciences. Therefore, the notion of measurement accuracy appears to cross-cut disciplinary boundaries. On the other hand, these studies raise questions about the feasibility and usefulness of using the methods of natural scientists when measuring parameters in the social and behavioural sciences. For instance, the measurement of national inequality admits different degrees of approximation, because scientists are willing to rely on slightly less accurate outcomes if this facilitates comparisons across countries and over time. What counts as the best epistemic strategy for the purpose of informing redistributive policies within a country is not the same as that for making comparisons across countries and time periods. Therefore measurement assessment strategies must be tailored to the purpose at hand. To deal with context-dependency, moreover, social and behavioural scientists have developed flexible measurement procedures, which allow them to trace regularities in context-dependence and thereby provide grounds for comparisons across contexts. For example, the measurement of relative poverty can be adapted to different living standards across countries. In my view, therefore, approximation and context-dependence are not problems as such, but rather guides measurement methodology and the employment of the outcomes.

This introduction surveys some central debates in the epistemology of measurement, especially as related to social and behavioural sciences, after which it highlights the main questions that are addressed in the dissertation articles.

2. The epistemology of measurement

The contemporary philosophical investigations on the epistemology of measurement form a broad and variegated literature that engages with a variety of topics concerning measurement practice and the tasks it involves. In this section, I focus on three debates in particular, all of which have far-reaching implications across disciplines.

The studies included in this dissertation adapt and combine insights from these debates to address the challenges that arise in social and behavioural measurement.

2.1. The problem of coordination

Scientific theories and models contain quantitative parameters like distance, temperature, volume, growth rate, and unemployment. To achieve empirical significance, these parameters must be linked, or 'coordinated', with procedures that enable us to determine their values. Philosophers have long been interested in how measurement allows theoretical concepts to be linked to empirical indicants. The problem is that, prior to the construction of an accepted measurement procedure, there is no evidence to confirm the rule for are assigning values to the parameter of interest. For instance, background theory concerning temperature and its relationship to the expansion of thermometric substances is required for the design of a thermometer. Testing this theory necessitates a reliable method for measuring temperature; yet checking the thermometer's reliability presupposes the same background theory that one wishes to confirm (cf. Collins 1985). Thus whether the changes in volume of a certain substance are quantitatively proportional to changes in temperature remain underdetermined (Chang 2004, Ch. 2).

Conventionalists about measurement have attempted to break the circle by stipulating a priori definitions that link parameters with specific measurement procedures (Mach 1896; Reichenbach 1927). This view emphasises that measurement involves non-trivial choices among alternative principles of coordination, instrument design, and measurement execution. Since there are no evidential grounds to adjudicate between these alternatives, conventionalists maintain that scientists converge on agreement, and choose a set of rules to standardise the measurement practice. In other words, coordination is obtained by agreed-upon definitions: it is an act of stipulation. The problem with this solution is that the choices lack justification, and it does not provide a criterion for improving those choices. Alternative measurements, which could possibly give conflicting results, would seem equally justified. In measurement practice, in contrast, procedures are typically chosen on the basis of empirical considerations and are sometimes replaced with others that are deemed more accurate.

Recent philosophical works, most notably including those of Hasok Chang (2004) and Bas van Fraassen (2008), have provided a new way of looking at coordination, by taking a historical and coherentist approach to the problem. Instead of trying to avoid the circularity, they show that it is not vicious because the definitions of the parameters and their measurements co-evolve. This interpretation highlights that knowing what is being measured and how to measure it are not independent tasks. To use van Fraassen's words, the questions 'What is a certain quantity?' and 'What counts as a measurement of that quantity?' cannot be answered independently of each other (van Fraassen 2008, p. 116). The historical development of measurement, however, allows progress to be made in both these tasks. Defining a parameter and designing a measurement procedure of that parameter are mutually dependent tasks that proceed progressively by iteration. According to Chang, the historical progress of thermometry involved back-and-forth relations between empirical interventions and theoretical developments. On the one hand, the

construction and testing of thermometers required underlying theoretical assumptions that could only be provisional (e.g., the linear expansion of thermometric substances) or more widely accepted thanks to confirmation coming from other fields (e.g., the law of thermal expansion of gases). On the other hand, the empirical evidence gathered led scientists to amend and refine theory and its concepts, e.g. to cast doubt on the assumption of linear expansion of thermometric substances. Each step improves on the previous conceptualisation of the quantity and its measurement procedure and refines the coherence among them. In this process, coordination is successful because it increases coherence between abstract definition and measurement procedure.

According to van Fraassen, we can see how this process avoids vicious circularity by looking at it 'from above', that is, in retrospect given our current knowledge about already stable and established measurements, or 'from within', that is, by looking at the historical stages where other procedures can be taken as given. It is only if one tries to take a 'view from nowhere' and attempts to find a coordination free of presuppositions or previous theoretical commitments that the process erroneously appears to lack epistemic justification (van Fraassen 2008, p. 122). On this view, measurement procedure and background theory co-evolve by mutually refining each other.

In this dissertation, I argue that the redefinitions of the parameter and of the way to measure it are part and parcel of measurement practice across the sciences. The measurement of social and behavioural phenomena often provokes heated debates both about what is being measured and about how to measure it. This, however, is a common problem with all new measurements, which are not yet established in scientific practice. In article II, moreover, I argue that the mutual refinement of definitional and procedural aspects is a characteristic of classificatory processes too. To defend this claim, I introduce a distinction between *classification systems*, which refer to how phenomena are ideally split and lumped according to an underlying organising principle, and *classifying methods*, which indicate the concrete procedure for assigning single cases to classes. By looking at recent research efforts into the classification of psychiatric disorders, the article argues that progress in classification arises from the mutual development of classification systems and classifying methods.

2.2. Model-based accounts of measurement

Model-based accounts further develop the idea that measurement involves interdependent theoretical and procedural aspects by emphasising the relation between measurement and theoretical and statistical models. According to model-based accounts, measurement involves two interrelated levels: a concrete procedure where an instrument interacts with the targeted object and the surrounding environment; and a model of that process, constructed from simplifying assumptions. On this view, measurement proceeds by representing the concrete interactions with model parameters and assigns values to those parameters based on the observed results of these interactions (Tal 2020).

Model-based accounts distinguish between instrument *indications*, such as the position of the fluid inside a thermometer or the location of the pointer on a balance, and measurement *outcomes*, which are knowledge claims about the quantity being

measured, like 'the temperature of x is 23°C' and 'the weight of y is 1.2 kg'. According to these accounts, measurement involves making inferences from indications to outcomes, and these inferences depend on model assumptions about the object being measured, the instrument, and the environment. Indeed, scientists can use the same procedure and indications to measure different parameters depending on how the measurement process is represented (Mari 2003; Tal 2019). Therefore, measurement presupposes a representation of the measurement process, that is, a model that represents the quantity under measurement and the measurement instrument in the ideal situation in which interfering factors are absent or controlled for.

The model has two main functions: it is needed for making inferences from indications to outcomes, and it provides the necessary context for evaluating the accuracy of these inferences. First, by making assumptions about how the measurement process would work in isolation from all interfering factors, the model justifies how measurement can assign values to what is measured. This dissertation emphasises that the definition of parameters in the natural, social and behavioural sciences contain assumptions that can only be approximated by the procedures for measuring these parameters. As a consequence, scientists across the sciences deal with a similar problem of coordination between theoretical parameter and measurement procedure.

Second, the model provides the theoretical basis to detect interfering factors and possibly control for their effects. Tal (2019) notes that instrument indications are subject to the idiosyncrasies of the concrete measurement process, such as interference from the environment, the operator, and the particular features of the instrument. For example, in the measurement of temperature, indications are influenced by changes in the volume of the thermometer glass. Measurement outcomes, instead, are expected to be invariant to these interferences. By looking at indications alone, one cannot distinguish between variations that are due to the influence of interfering factors and variations due to changes in the quantity under measurement. According to model-based accounts, measurement errors are evaluated by the distance between a given outcome and the value that would be expected in the absence of interfering factors (Tal 2019, p. 871). Therefore, it is only by reference to the ideal measurement conditions depicted by the model that scientists can evaluate and possibly correct for the relevant sources of error. On this view, therefore, the accuracy of measurement depends on the procedure as well as on how it is represented.

Model-based accounts of measurement provide a notion of measurement accuracy that is consistent with the evaluation of uncertainty and the correction of systematic errors in metrology, the science of measurement (JCGM 2012, 2.13 Note 3; Giordani and Mari 2013; Tal 2011; 2017). On these accounts, accuracy is a metaphysically neutral concept that does not imply the existence of true values. Epistemic accuracy is the term used to describe this concept (Tal 2011: 1084-5). As I discuss in article **I**, the epistemic accuracy of measurement can be evaluated by testing the robustness of multiple outcomes of the same (or related) quantity, given the different ways in which they are modelled (Staley 2020). When compared to the evaluation of measurement in traditional accounts of measurement, like realism and conventionalism, this notion of accuracy becomes clearer.

2.3. Evaluating the accuracy of measurement

According to a realist, error-based perspective, the outcomes of measurement can be evaluated in terms of their distance to the true values of the parameters being measured (Swoyer 1987). On this view, measurement outcomes are *accurate* if they are close to the true value of the parameter, and *precise* if they are close to each other. Accuracy and precision are typically described using the image of an archery target, where the bull's eye is analogous to the true value of the measured quantity. A measurement is precise if the arrows land close to each other, and accurate if they get close to the centre of the target (JCGM 2012). Ideally, a measurement should aim at maximising both accuracy and precision, so as to get all arrows close to the target centre. But a measurement can have high precision and low accuracy (if the arrows are tightly grouped, but not quite at the centre of the target). Precision is a necessary but not sufficient condition for accuracy: a highly precise measurement is not guaranteed to be a good measurement of the parameter of interest, because it could be affected by a systematic source of error while remaining consistent under repetition (Zeller and Carmines 1980).

Realism about measurement appeals to a metaphysical notion of accuracy, which is problematic because we have no access to true values. Despite its clarity, the error-based perspective has limited practical applicability, because the true values of most quantities of interest are unknowable or undefined, even assuming there are such things. Thus accuracy, so conceived, also remains underdetermined by evidence. It can be estimated by comparing fallible measurements to one another, yet it is unclear whether the convergence of fallible measurements can be taken as proof that the common outcomes are true: they could all be plagued by a shared systematic error (Cartwright 1991).

Article II argues that contemporary debates concerning psychiatric classification are framed by the opposition between alternative notions of measurement accuracy. Modern, symptom-based classifications rely on a conventionalist notion of accuracy, because they postulate clear and unambiguous indications for diagnosing mental disorders that clinician of different theoretical orientations can agree on. In contrast, the proponents of causal classifications advocate a realist notion of accuracy, which prioritises the discovery of the real causal mechanisms behind mental phenomena. In this article, I argue that both approaches are problematic and do not provide tools for improving psychiatric classification. Instead, I suggest that improving psychiatric classification is a feasible goal if we adopt an epistemic notion of accuracy.

In the absence of epistemic access to true values, scientists have developed strategies to detect, distribute and correct errors by comparing outcomes to each other. As I argue in article I, the comparison of different measurements of the same parameter is central to measurement assessment practice, and it is common across the natural, social and behavioural sciences. The consistency of outcomes across different measurements of the same parameter is meant to ensure that the outcomes are about what is measured, rather than being influenced by artefacts of the instrument, the environment, or the model. Tal (2019) emphasises that this comparison is meaningful only if the measurements are modelled in terms of the same quantity of interest. Conditional on this judgment, agreement can be taken as a sign of accuracy, and discrepancies can be interpreted as pointing to undetected errors: by comparing

measurement outcomes in the light of their respective models, scientists can detect errors due to various interfering factors. This involves both manipulating the model and intervening in the process: improving the epistemic accuracy of measurement might require altering the process, but it might also be done without physical interventions, by adjusting or modifying the representation. Measurements are deemed accurate if their outcomes agree within their respective uncertainty intervals (Tal 2011).

Continuing with the example from the history of thermometry, scientists detected the interfering effect of the different rates of expansions of glasses by observing the disagreement between thermometers. Once the outcomes are corrected to account for this interfering factor, the thermometers are made to agree without any physical interventions on the instrument, the object under measurement or the environment. Similarly, I argue that a questionnaire and an interview for diagnosing migraine can be found to disagree because the former, but not the latter, tends to underestimate the duration of headache due to the subjects taking medication or sleeping through it (Article **I**). After the error has been corrected, the two diagnostic procedures are made to agree within their respective levels of uncertainty. Since the correction of systematic errors can raise uncertainty, accuracy improves when the errors are corrected with a minor increase of uncertainty (Tal 2019). Discrepancies in measurement outcomes can be regarded as pointing to an undetected error, but they could also be interpreted as indications that the instruments measure different quantities. According to Tal, this is because the detection of systematic errors and the individuation of quantities are two sides of the same coin. Scientists can choose between the two interpretations based on the historical and theoretical development of the measurement (Tal 2019).

3. Measurement in the social and behavioural sciences

Social and behavioural scientists measure things like welfare, well-being, utility, poverty, inequality, customer satisfaction, consumer price index, unemployment, quality of life, learning, and performance. Measurement is also used to investigate personality traits and mental disorders, as well as the prevalence of conditions or events, like alcohol abuse, crime, early school leaving, etc.

One of the key roles of measurement in these fields is to monitor change over time and make comparisons across regions and contexts. Another key role is to enable the exploration of relations among parameters. Claims like "we have experienced increasing inequality and global poverty", "public education systems reduce inequality among children", or "the mortality rate for psychiatric diseases has not fallen in the last 50 years" are based on systematic measurements of the parameters involved. In this way, measurement contributes to uncovering comorbidities, patterns of behaviours, and causal relations between variables.

We can only see these changes and these relations because standardised measurement has been used. As mentioned before, measurement is thought to allow the consistent and systematic investigation of a phenomenon across contexts and over time. This is based on the idea that the changes in the outcomes are due to changes in what is measured, rather than to the idiosyncrasies of the instrument, method, procedure, or surrounding environment. Therefore, whether and how social and

behavioural scientists can succeed in providing consistent and standardised measurement across contexts is a crucial issue that bears heavily on the policy and practical implications of measurement in these fields.

The reliability of measurement in the social and behavioural sciences (that is, its ability to provide knowledge for what is measured), however, is often contested. First, the measurability of social and behavioural phenomena is a topic of continual debate (e.g., Michell 1999; Moscati 2018; Alexandrova 2017). The complexity of the phenomena involved and the influence of intentional action and personal mediation make it difficult to come up with precise definitions of the parameters of interest. Moreover, many of the parameters measured in these fields are context-sensitive and this is an obstacle to developing standardised procedures that can be used across contexts. Third, the reliability of these measurements is questionable because of the difficulties of obtaining the kinds of experimental control that can be achieved in the laboratory. Finally, while measurement is commonly lauded for introducing an automatic or mechanical procedure that substitutes subjective judgment and is value-free, this is not always possible, especially outside the physical realm. Epistemic and ethical considerations are often deeply intertwined in social and behavioural measurement, and this impacts their reliability because the outcomes can be contested on ethical grounds.

In the remaining of this section, I illustrate these challenges in greater detail and highlight the questions addressed in this volume as well as those that remain open for further investigation.

3.1. Parameter definitions and their measurement

The recent works in the epistemology of measurement discussed above emphasise that the definition of a parameter and its measurement are interdependent task. Therefore, in this view, the conditions for measurability lie not solely in the characteristics of the phenomenon itself, but rather depend on how the phenomenon is defined within the relevant theory, and on how this definition is linked to empirical indicators. The measurability of a parameter depends on the coherence between the definition and the procedure. Drawing on this insight, Cartwright and colleagues stress the importance of definitional concerns in the social and behavioural sciences by arguing that measurement in these fields requires alignment between the definition of the parameter of interest and the relevant measurement procedure. More precisely, on the one hand, the procedure should capture all and only the dimensions of the parameter as defined, and, on the other hand, the definition cannot include dimensions that the procedure is unable to measure (Chang and Cartwright 2008; Cartwright, Bradburn, and Fuller 2017). For example, a clinical procedure for diagnosing schizophrenia should consider all the aspects included in the relevant definition of schizophrenia such as delusional perception, persistent hallucinations, breaks in the train of though, and incoherent speech. Factors that are irrelevant to this definition, like the subject's socio-economic status, should not influence the diagnosis. On the other hand, factors that cannot be captured by the current clinical diagnostic methods, like structural brain changes, cannot enter the definition of schizophrenia for clinical diagnostic purposes, but could instead be relevant for research purposes (e.g., the study of treatment efficacy).

In the social sciences, but also in psychology and psychiatry, however, the definitions of the parameters under measurement are rarely as widely accepted as those of well-defined physical quantities. Some parameters have official definitions, because there are institutions that identify and recommend highly reproducible measurement procedures to be employed in several contexts, thereby creating a common language for providing comparable results. For instance, the American Psychiatric Association publishes the Diagnostic and Statistical Manual of Mental Disorders, which provides the definitions of mental disorders that are relied upon not only by clinicians and researchers, but also by health insurance companies and policy makers. Similarly, in economics, the System of National Accounts provides a set of international recommendations for the measurement of parameters like national gross output, income, social spending, and household consumption. However, even with respect to these officially recognised definitions, the consensus is far from unquestionable.

Indeed, in the context of social and behavioural measurement, coming up with a precise definition of the phenomenon of interest can face specific problems. According to Hand (2004), one important reason for these controversies is the sheer complexity of the behavioural sciences: in psychology, virtually all variables are related to each other, and it is difficult to tease the complex tangle apart. Mental phenomena, moreover, are mediated by the individual and, while in physics all electrons are identical, in psychology no two people are the same (Hand 2004, p. 152). In economics and other social sciences, instead, many of the parameters of interest are aggregate objects, that is, statistical constructs that aggregate individual objects. This raises concerns related to the existence of these higher-order objects and how aggregation methods can influence the measurement outcomes and their interpretation (Hand 2004; Desrosières 1998, p:70; Porter 1995).

Numerous authors have emphasised that the parameters measured in the social and behavioural sciences are defined differently across contexts (Cartwright and Runhardt 2014; Alexandrova 2008; McClimans 2010; Angner 2013). Cartwright and colleagues put forward the notion of *Ballung* concepts to indicate that many concepts in the social and behavioural sciences are multifaceted and loosely defined (Cartwright, Bradburn, and Fuller 2017; Cartwright and Runhardt 2014). Concepts like poverty, inequality, depression, and well-being are characterised by clusters of features and have no central core without which one does not merit the label. Different sets of features can matter for different uses: which features are relevant depends on the context and purpose. These concepts are bound to have multiple meanings, each of which is likely to sacrifice or alter some aspects of the concept to emphasise others. For instance, scientists can use a variety of indications to measure 'cultural decline', such as reduction in welfare, violent crime, suicide, and early school leaving. But it is immediately evident that disagreement is to be expected about which indicators the definition should include (Hand 2004). Therefore claims like "we have experienced cultural decline" can be contested on definitional grounds.

The measurement of these multifaceted concepts can raise two related problems. First, the investigation of *Ballung* concepts can produce a proliferation of heterogeneous measurements. Measuring these concepts often requires trading off two conflicting desiderata: tailoring the measurement to fit the intended purpose and being able to compare measurements with each other (Cartwright, Bradburn, and

Fuller 2017). For instance, as I discuss in article II, national inequality measurements are usually tailored to fit the purposes of the country where the measurement is carried out and, as a result, national inequality measurements vary greatly across countries and over time. The lack of uniformity among national inequality measurements, however, creates problems in investigating the differences across countries and trends over time: one cannot be sure whether the results of comparisons are genuine or artefacts of measurement differences. For making comparisons and investigating trends over time, it would be better to have a standardised measurement, which can be used across contexts and over time. Standardised measurements, however, are less able to capture context-specific features.

The second problem concerns the scope of application of these heterogeneous measurements. Broad and multidimensional concepts are difficult to measure, because no procedure can take into account all their relevant features at the same time. Purpose-specific measurements, instead, typically rely on narrower concepts, which are more tractable and measurable. However, one can wonder what the relation is between these precisely defined but narrow parameters and the broader concepts that are relevant to some practical purposes and policymaking. Consider this example about inequality measurement.

In policy-making, inequality has been defined as "the fundamental disparity that permits one individual certain material choices, while denying another individual those very same choices" (McKay 2002, p.1; Ray 1998). But measurement requires a different definition, which allows this parameter to be coordinated with empirical indicators. By looking at contemporary measurement practice, McGregor et al (2019) formulate the following definition of inequality: "a property of a variable's frequency distribution within a population, which is typically summarised in a single statistic". When measuring inequality, scientists narrow down this definition by choosing the resource and population of interest, and by using a particular statistical index. For instance, scientists can measure income inequality across Europe with Gini coefficients, or consumption inequality in Spain with a different statistical index. This results in a proliferation of heterogeneous inequality measurements. The problem with these narrow definitions, however, is that they are devoid of what makes inequality interesting in the first place, that is, its potentially negative consequences for individuals and societies. So the question remains of whether these narrow measurements are also relevant to addressing broader issues, for instance the effect of globalisation on international inequality. How are these narrow measurements related to each other and to the broader, multidimensional concept?

To address these questions, this dissertation looks at cases where scientists seek to make comparisons across context-specific measurements of the same broader concept. Overall, I defend the idea that, when investigating context-dependent phenomena, it might be possible to tailor the measurement to context-specific features and still be able to compare across partially heterogeneous measurements, with some degrees of approximation.

A well-rehearsed example is discussed in article I. In the measurement of poverty as relative deprivation, the definition of poverty is context-sensitive because it depends on the living standards of the society under consideration. However, the procedure for measuring relative deprivation retains some standard features across contexts. While the list of basic living conditions changes from country to country,

each list is based on an underlying threshold that remains fixed: it aims to capture what is essential for a decent living condition, whatever this might be in the relevant society. Therefore, there is a sense in which these contextual definitions are all related to a broader concept of poverty, and this allows scientists to make comparisons across contexts. In other words, measurement is tailored to the context of application, but also retains some standard features that make it comparable across (partially) different methodologies. This example suggests that, in some cases, a flexible procedure can be used to measure context-dependent phenomena.

Article II further discusses this issue by looking at how scientists deal with the problem of comparing heterogeneous measurements of national inequality. Unlike the measurement of relative poverty, in this case there is no standardised procedure that can be adapted to different contexts. Instead, scientists have developed strategies to improve the comparability of context-dependent measurements by harmonising their outcomes. In article II, I argue that harmonisation worsen the accuracy of the outcomes. I suggest, however, that scientists can improve these strategies by changing the assumptions on which they rely. Moreover, I warn that, because of context-dependency, not all measurements can be harmonised without compromising the meaningfulness of their outcomes.

A question that must be further investigated is whether the cross-context comparisons can be justified on the basis of higher-order conceptual analogies, and what the implications of such analogies are for measurement and its multiple applications.

3.2. Measurement assessment practice and the notion of accuracy

The recent philosophical literature on measurement has emphasised that, in the absence of independent epistemic access to the quantity being measured, scientists evaluate their measurements by comparing fallible measurements to each other. In this dissertation, I argue that this strategy is used across the natural, social, and behavioural sciences: scientists face similar challenges and could benefit from similar ways to address them. But there are differences too.

In the measurement of well-defined physical quantities, this comparison amounts to controlled variations of the measurement instrument, because the instruments are modelled to a high level of detail. In the measurement of time, for instance, different atomic clocks realising the standard second differ in the way and the degree to which they approximate the ideal conditions of the model. For example, different atomic clocks approximate the assumption of absolute zero temperature in different ways and with different degrees of associated uncertainty.

Social and behavioural scientists, however, might not have access to this strategy in the same way as the measurement of well-defined physical quantities. On the one hand, social and behavioural measurement might be unable to meet the conditions for repeated observations, which include adopting the same measurement procedure, the same observer, under the same conditions and in the same location over a short period of time (Boumans 2015). The quantity being measured might be known to change over time; measurement might be costly and time consuming (e.g., in the case of large-scale surveys). Measurement itself might be thought to influence the quantity being measured (Hacking 1995). On the other hand, the models of social

and behavioural measurements might not allow for the same kind of controlled variations, because they do not model the apparatus to the same level of detail, or because they rely largely on unrealistic assumptions and ceteris paribus conditions. As a consequence, it might be more difficult to predict the effects of each single way in which the measurement procedure departs from the ideal model (Boumans 2006; 2015).

Simplifying assumptions are also required in the measurement of physical quantities. However, in successful examples of physical measurement, such as the realisation of the standard units of measurement, scientists evaluate every known way in which the model differs from the actual measurement process. The pervasive use of unrealistic assumptions in social and behavioural measurement is one of the reasons why, according to Boumans (2015), these measurements should not be tested based on how each assumption represents the target system, but rather on more general patterns of behaviour. For example, behaviour pattern tests look at how well the model as a whole can mimic the primary behaviour patterns found in the real system rather than testing the individual assumptions.

In this dissertation, however, I argue that social and behavioural scientists do test single assumptions, even if perhaps not all of them. Some corrections are performed routinely, such as when testing the sampling assumptions underlying survey measurements. Other corrections instead require ad hoc procedures. For instance, I discuss how economists detect and correct for the underreporting of top incomes in the measurement of national inequality. Survey data is likely to underestimate inequality because the very rich rarely participate in surveys and because extreme incomes are sometimes top-coded or eliminated as outliers. To estimate this source of error and correct the outcomes, scientists compare survey results to tax data. Although tax data still suffers from misreporting, especially at the bottom of the distribution, it is more reliable than surveys for top incomes. Therefore, economists can use the gap between tax and survey data to estimate the amount of underreporting of top incomes. As discussed in article **I**, a similar strategy can also be observed in the measurement of migraine prevalence, where scientists compare questionnaires and interviews to detect sources of error that affect the former, but not the latter measurement procedure. However, some of these corrections remain controversial, and not everyone agrees as to whether and how they should be performed. Therefore, it is still an open question as to how the overall reliability of these measurements depends on the extent to which the main sources of error have been tested.

3.3. Value-laden measurement

In the social and behavioural sciences, ethical considerations have an impact on the definition of some parameters under measurement, like poverty, inequality, and well-being. For instance, defining income inequality requires making value judgments about how desirable equality is at various points of the distribution. Similarly, a definition of subjective well-being (that is, how good a person's life is from this person's perspective) necessarily appeals to normative claims about what constitutes a good life.

Alexandrova (2008) emphasises the context-dependent nature of these value judgments: which ethical claim is better suited for defining the parameter of interest depends on the purpose that the measurement is meant to achieve. For instance, whether well-being has to do primarily with the absence of pain in the short run, or with success in one's life goals, depends on the context. A balance of pleasure and pain could be appropriate for the evaluation of certain medical interventions, but not for the evaluation of career choices. In other words, answering the question 'What is the impact of *X* on a person's subjective well-being?' requires different measurements depending on whether *X* is a medical intervention or a career choice.

Similarly, in economics, the measurement of discounting parameters has different ethical implications across contexts of employment. Time discounting refers to the common practice of weighting costs and benefits depending on the time at which they occur, so that the outcomes occurring far in the future are given less weight than more immediate ones. In most of its applications, how much to discount is a matter of convention, and the parameters are measured on the basis of market interest rates. As discussed in article IV, however, when it comes to choosing the discounting rates to be used in the evaluation of climate policies, ethical considerations of intergenerational justice have an impact on these parameters, because time discounting might lead to minimising the worse consequences of climate change just because they are likely to be experienced by future people living centuries after us.

Consequently, it appears that the reliability of these measurements, that is, whether they are apt for their purposes, hinges not only on epistemic considerations (such as reproducibility of the measurement and uniformity of the procedure) but also on these ethical and normative claims (see also McClimans 2010).

Based on this idea, Alexandrova's works note that scientists measure relatively narrow concepts of well-being, like the well-being of patients with a chronic disease or the well-being of students from foster homes (Alexandrova 2008; 2016; 2017). In contrast, broad conceptions of well-being, such as those that call for a comprehensive aggregation of people's preferences, are not measurable, because no measurement could take all its features suitably into account. Angner (2013) has emphasised that these narrow measurements are easily misinterpreted: scientists might uncritically use the available, narrow measurements in situations where broader concepts are implied, giving rise to equivocation fallacy and reification.

This problem is associated with particular uses of these measurements that exaggerate their scope and content. It is not a problem for narrow measurements per se, as far as their employment remains within their scope and meaningfulness. However, this leaves us with the open question of what to do when we are interested in broad and morally charged parameters whose ethical implications go well beyond what can be measured with questionnaires and the observation of market behaviour. Article IV addresses this question in the context of the discounting parameters used in the evaluation of climate policies. In this context, the discounting parameters hang on ethical judgments that go well beyond what can be captured by market-based values, and some economists have therefore begun to treat these parameters as reflecting ethical principles rather than solely as quantities to be measured. Choosing the value of discounting rates on the basis of ethical considerations, however, is not a straightforward matter. Therefore, evaluating the role of ethical considerations in the

measurement of morally charged parameters and their effects on measurement accuracy remains a matter for further investigation.

4. Research methodology

The studies of this dissertation rely heavily on case studies to illuminate scientific practice and its epistemological presuppositions in various social and behavioural sciences: economics, sociology, psychiatry, and medicine. The case-based approach reflects the belief that epistemological concerns like justification, observation, and reliability are best investigated in close contact with actual scientific practice (Ankeny et al. 2011; Bursten 2020). While this requires thorough engagement with the relevant scientific literature, the articles remain eminently philosophical. Accordingly, the dissertation does not provide new empirical findings; instead, the methods employed are conceptual analysis (broadly conceived) and critical reflection. The four articles provide detailed and systematic analysis of scientific activities and explore their epistemological presuppositions. More precisely, the methodology of this dissertation can be described in four steps:

- clarification of what kind of activities are involved in the generation of knowledge in social and behavioural measurement;
- investigation of the assumptions and inferential patterns underlying these activities;
- critical reflection on how these assumptions and inferences shape the knowledge claims produced by measurement; and
- exploration of the implications of how such knowledge is used in science and policy.

This methodology has the potential to produce results that are relevant for scientists as well. In other words, these studies are not only about illuminating scientific practice and its epistemological presuppositions, but can as well contribute to improving this practice. In particular, they can promote a thoughtful approach to measurement practice and ways of employing the outcomes that are conscious of the underlying measurement assumptions and their implications.

5. Overview of the articles

Article I delves right into one of the central tasks of measurement practice: the assessment of measurement. By looking at examples like the measurement of time, the measurement of poverty, and the measurement of migraine prevalence, this article argues that scientists use similar strategies across disciplines, and it outlines the inferential pattern behind these common strategies. Articles II and III draw on these insights to address specific contexts: the measurement of inequality and the classification of mental disorders, respectively. Article II addresses the problem of comparing heterogeneous measurements of national inequality, which are based on different methods and presuppositions. This article suggests a way to improve the quality of these comparisons and outlines the conditions under which such comparisons are legitimate. In article III, insights from contemporary philosophy of

measurement are adapted to interpret the debate about the classification of mental disorders and this helps to address the question of how this classification can be improved in the absence of cognitive access to real taxonomies (even assuming there are such things). Finally, article IV addresses the challenges of value-laden measurement practice by investigating the choice of the social discount rate to be used in the cost-benefit analysis of climate policies. The article argues that, given the prominence of ethical considerations bearing on this parameter, the choice should be based on ethical principles, and it explores the feasibility of doing so.

5.1 The appeal to robustness in measurement practice

Article I looks at how scientists appeal to robustness in evaluating the accuracy of measurement and provides a reconstruction of the underlying argumentative pattern. This assessment strategy, I suggest, presents important similarities across disciplines. The article looks in particular into the measurement of time, of the endowment effect, of migraine prevalence, and of poverty. Across all these contexts, scientists deal with a similar problem of coordination between the theoretical definition of the parameter of interest and the actual measurement procedure. The definitions of the standard second, of migraine, and of poverty all contain assumptions that can only be approximated by the procedures employed to measure these parameters. For instance, the definition of the standard second makes reference to an unperturbed caesium atom, that is, it assumes that the atom is at absolute zero temperature, under conditions where there is no gravity, no magnetic field, etc. Similarly, the definition of migraine prevalence assumes that patients can correctly evaluate the duration of their headaches, and some definitions of material poverty assume that individuals have similar basic needs across countries and time periods. Alternative procedures approximate these assumptions in different ways and to varying degrees, and therefore have different sources of uncertainty. For instance, depending on their specific constructing features and surrounding environment, different atomic clocks realise the condition of absolute zero temperature with different degrees of approximation. Similarly, in the measurement of migraine prevalence, questionnaires are more exposed than interviews to the underestimation of headache duration due to the subjects taking medications or sleeping through the headaches. In fact, the physician can ask additional clarifying questions during the interview.

It is thanks to these differences that, by checking the robustness among alternative procedures, scientists can evaluate and improve their measurements. This is how it works. Based on how each procedure departs from the definition of the parameter being measured, scientists formulate hypotheses concerning how this affects the outcomes and, if possible, introduce a correction to the results. For instance, if the actual temperature of the atomic clock is expected to stably influence its ticking frequency, scientists can correct its outcomes by the expected influencing factor. The accuracy of measurement improves if this correction increases the convergence of the outcomes with those of other procedures. This means that scientists have minimised the uncertainty due to the differences between procedures and tightened the coordination with the theoretical definition.

This assessment strategy is different from common robustness analysis because it is based on a different argument. Robustness analysis is usually interpreted as based

on a no-coincidence argument: it would be surprising that different procedures converged on the same outcomes if they were not correct. The appeal to robustness in measurement assessment instead looks at how the procedures should converge given the different ways in which they approximate the parameter being measured. It is these expectations that are tested by comparing fallible procedures to each other, rather than the outcomes themselves.

This assessment strategy provides confirmation for a particular notion of measurement accuracy, which depends on the coherence between the definition of the parameter and the measurement procedure. In other words, the appeal to robustness in the assessment of measurement provides an indication of how well the procedure captures what it is supposed to measure, as defined. Since this inferential pattern recurs across the natural social and behavioural disciplines, this notion of measurement accuracy appears to be relevant across disciplinary contexts. Therefore, this study suggests that there is continuity across the natural, social and behavioural sciences in the notion of measurement accuracy and its evaluation methods.

5.2 Measuring inequality across countries and over time

Comparing inequality across countries and time periods is difficult because national inequality measurements are based on country-specific methods and presuppositions, which also vary over time within the same country. In the absence of an agreed-upon methodology, each country makes its own decisions, resulting in a patchwork of indicators that are not necessarily comparable, are difficult to aggregate, and might tell contradictory stories. For instance, when comparing inequalities based on net and gross income, it is difficult to disentangle the effect of redistributive taxation.

Researchers seeking cross-national comparisons of inequality rely on harmonising methods that allow them to express heterogeneous measurements in homogeneous terms. For instance, if one finds that inequality outcomes based on gross income are on average 5% higher than those based on net income, one could use this additive adjustment factor when one parameter is available but not the other for a specific country and year. These methods, however, have been criticised by showing that they rely on strong assumptions and might lead to misleading conclusions.

Article II argues that harmonising methods can be interpreted as ex-post corrections to the measurement outcomes that do not require intervening in the actual process: economists harmonise the parameter under measurement and transform the outcomes accordingly, without running a new survey. Based on this interpretation, the article defends two claims, one methodological and the other substantive. The methodological claim is that harmonisations can be improved by mitigating their detrimental effect on measurement accuracy. The substantive claim is that harmonisations can be legitimate under the condition that they do not compromise the meaningfulness of the outcomes.

Commonly used harmonisations have a detrimental effect on the accuracy of measurement because they are based on implausible assumptions that add new sources of uncertainty. For instance, harmonisations are based on the assumption that the correlation between the source and the target parameter is constant. In order to harmonise outcomes based on different definitions of income, such as gross and net

income, the procedure assumes that the relation between inequality measured using one income definition and inequality measured using another is constant across time and between countries. This assumption is implausible for a number or reasons, for instance because the redistributive impact of taxation varies across countries and time. Thus harmonisations based on this assumption worsen the accuracy of measurement, because the different redistributive effects of taxation across countries is not controlled for. This suggests that harmonising methods can be improved by modifying the assumptions on which they are based to account for country-specific characteristics. Rather than relying on a constant adjustment factor, harmonisations that take country-specific details into account might be less detrimental on accuracy.

Harmonisations, moreover, appear to present questions of legitimacy when seen in this light. When economists harmonise, they in fact alter the heterogeneous definitions of inequality rather than running a new measurement, and this might affect the interpretation of the harmonised outcomes. For instance, in a country where there is a gender gap in property rights, the choices of the reference units are crucial to representing the distribution of income among the population and cannot be changed without creating distortions. A mechanical harmonisation of the reference unit would be inappropriate, and produce a distorted picture of the income distribution among the population, besides being conceptually wrong. Harmonisations, therefore, can only be legitimate when they alter the original measurements without compromising the meaningfulness of their outcomes.

5.3 From measurement to classificatory practice: improving psychiatric classification independently of the opposition between symptom-based and causal approaches.

Article III draws on insights from contemporary philosophy of measurement to propose a new way of looking at classifications in general and psychiatric classification in particular. This perspective, I suggest, has particular merit in the context of psychiatric classification, because it helps us to see how classifications can be improved independently of the stiff opposition between causal and symptom-based approaches. The premise of this work is that classification is a form of measurement in which the outcomes are nominal rather than quantitative or ordinal. This is an acceptable premise within a model-based account of measurement.

From a measurement perspective, symptom-based classifications appear to be based on a conventionalist view of measurement, which focuses on standardising classificatory practice so that different subjects can agree on the same definitions. In contrast, causal classifications reflect a realist view of measurement, which prioritises the aim of capturing the real causal factors of mental disorders. Both accounts face problems in the evaluation and improvement of classification. Symptom-based classification appears to rely on an operational notion of accuracy, which depends on compliance to agreed-upon definitions. This notion is problematic because it involves an element of conventionality and lacks a clear justification of why the agreed-upon definitions should be standard. Causal classifications, on the other hand, are based on a metaphysical notion of accuracy, which depends on having access to how mental phenomena are split and lumped independently of our classifications. The problem

with this notion is that true classifications are unknowable, even admitting that there are such things.

A measurement perspective suggests that improving the accuracy of classification is a feasible goal if we adopt the epistemic notion of accuracy derived from model-based accounts of measurement. This notion of accuracy depends on the coordination between theoretical definitions and classification procedures, and it is formulated by introducing a distinction between classification system and classifying method: the former refers to how phenomena are split and lumped ideally according to an underlying organising principle; the latter indicates the concrete procedure for assigning individual cases to classes. As theory and measurement coevolve, classification systems and classifying methods can also be part of a process of mutual refinement.

To support these ideas, the article argues that scientists can improve the epistemic accuracy of psychiatric classifications by comparing different classificatory methods to each other. This strategy is illustrated with an example. Scientists compare disciplinary perspectives on specific mental disorders, based on 'local' representations of the interaction between causal factors, rather than relying on a comprehensive model of the mental disorders' complex aetiology. In this process, classification systems and classifying methods coevolve by mutually refining each other. Since the success of this strategy does not depend entirely on whether mental disorders are initially classified in terms of symptoms or causes, the article suggests that the opposition between these alternative approaches is of little consequence in making progress in the epistemic accuracy of classification.

5.4 When utility maximisation is not enough: Sufficientarianism and the economics of climate change

Article IV examines how scientists choose a value for the Social Discount Rate to be used in the economic evaluation of climate policies. Crucial ethical considerations affect these choices, and this calls into question the measurability of the discounting parameters to be utilised in this context.

The economic evaluation of climate policies raises ethical concerns of intergenerational justice because today's choices will affect the well-being of future people. Scientists predict that future people living centuries after us will experience the most serious effects of climate change. The economic evaluation of climate policies, however, tends to give those later consequences much less weight than earlier (often lesser) effects, due to the common practice of discounting costs and benefits that occur in the future. The weighting of costs and benefits according to the time at which they occur is typically done on the basis of a parameter called Social Discount Rate (SDR), which allows us to estimate the present value of future costs and benefits. The evaluation of climate policies is highly sensitive to the SDR and hence the value assigned to this parameter can have a significant impact on today's choices. In the ethics and economics of climate change, however, how to assign a value to this parameter is a point of contention.

According to economic theory, the SDR depends on two other parameters, the marginal benefit of consumption (η) and people's time preferences (δ), which reflect certain characteristics of people's preferences. The parameter δ, for instance, reflects

the idea that a vacation in Honduras is more valuable to me if I can enjoy it right now than if I have to wait 5 years for it. Some economists claim that δ and η can be measured empirically, for instance with behavioural economic experiments or by observing people's market behaviour. However, when the SDR is applied in the evaluation of climate policies, these parameters have implications that go well beyond what can be captured by these measurements, especially with respect to their implications for intergenerational justice. Thus it would be misleading to assign a value to the SDR based on those measurements, because the parameters, as measured in the lab for instance, do not capture people's ethical principles concerning intergenerational justice. In this context, therefore, the choice of SDR requires a different method, and indeed some climate economists have treated these parameters as reflecting ethical principles rather than solely as quantities to be measured.

Choosing the value of δ and η on the basis of ethical arguments, however, is not a straightforward matter. Different ethical principles would justify different choices, and their application to economic analysis might require specific adjustments to the economic framework and methodology. Moreover, in some cases it is not obvious exactly which values are suggested by a given ethical theory. In this article, Simo Kyllönen and I consider a specific non-utilitarian approach to justice, *sufficientarianism*, which cares especially about people's basic subsistence. According to sufficientarianism, our primary moral concern should be to improve the situation of the people living below a basic threshold of decent life. We argue that sufficientarianism is better suited to guide the choice of the SDR than other ethical theories and discuss one possible way of doing so. Utilitarianism, for example, is unsuited to address these concerns, because it neglects issues of welfare distribution.

According to sufficientarianism, the current generation has an obligation to save for the future if future people are likely to fall below a basic threshold of decent life. We discuss which values are suggested by this principle and examine their implications. Choosing the SDR based on sufficientarian principles would require major changes in some of the core assumptions and customary methods of climate economics. For instance, while standard economic analysis assumes that all costs and benefits are mutually substitutable, for sufficientarianism the resources necessary for the satisfaction of people's fundamental interests cannot be substituted with others.

6. Concluding remarks

This dissertation studies measurement practice in the social and behavioural sciences and investigates the conditions under which measurement can be reliable in these fields. The articles included suggest that there is continuity across the natural, social, and behavioural sciences in the notion of measurement accuracy and its assessment methods. Scientists across these disciplines evaluate their measurements by establishing robustness among fallible measurements of the same or related parameters, based on their respective models. Since the procedures have different ways to approximate the definition of the parameter under measurement, looking at their robustness allows scientists to evaluate how well these procedures measure the parameter as defined. Under this interpretation, the accuracy of measurement is the closeness of agreement among multiple measurements of the same parameter.

Social and behavioural measurement, however, faces specific challenges that affect both the reliability of the outcomes and how they can meaningfully be employed. This dissertation emphasises that the measurement of social and behavioural phenomena involves greater degrees of approximation than the measurement of well-defined physical quantities. One of the reasons is that, in these fields, it can be difficult to meet the conditions for repeatability and to perform the kind of controlled variations that are typical in the assessment of physical measurement. Moreover, because of the extensive use of simplifying assumptions in social and behavioural models, scientists in these fields do not typically test all their underlying assumptions. The assessment of social and behavioural measurement, furthermore, appears to admit a certain degree of flexibility, depending on the purpose the measurement is meant to serve. How much flexibility a measurement can admit depends on its theoretical and procedural development.

Another set of challenges is related to context-dependency. Many of the parameters of interest in the social and behavioural sciences have context-dependent, morally laden definitions. In this dissertation, I address various reasons why measurement can be context-dependent, such as the different purposes the measurement is intended to achieve, the varying ways in which the phenomenon manifests itself across contexts, or the differing ethical implications that the measurement might present. Each of these creates challenges for the reliability of measurement. The most immediate problem with context-dependency is that it obstructs comparison and accumulation of knowledge: if the definition of the parameter depends on the context, it might be difficult to make cross-context comparisons. On the other hand, standardising the definition so that it can be used across contexts requires the foregoing of context-specific details and therefore can result in a loss of reliability. Because of context-dependency, moreover, the reliability of these measurements can be challenged on definitional or ethical grounds.

These problems, I suggest, can be mitigated. The studies in this dissertation argue that it is not enough to simply note that a measurement raises heated debates about what is being measured and how, because this is a common problem with all new measurements that are not yet established in scientific practice. To go deeper in evaluating the accuracy of these measurements, one must look at how they are assessed and improved. This constructive approach permeates each of the dissertation articles. This dissertation addresses the challenges of social and behavioural measurement by clarifying the impact of scientists' strategies on the accuracy and the meaningfulness of the outcomes. I argue that redefinitions of the parameters and their measurement procedures are not only common practice across the sciences, but they are also necessary for improving those measurements. Moreover, I suggest ways to improve the accuracy of measurement and warn against ways of using the outcomes that are conceptually and/or epistemologically inappropriate. In my view, therefore, context-dependency and approximation are not problems per se, but can rather be seen as guiding scientific practice and the scope of application of each measurement. So conceived, the study of the epistemological presuppositions of measurement practice contributes to building awareness of the conceptual and ethical implications of scientists' strategies in these fields, and promotes thoughtful ways of using the outcomes.

References

Alexandrova, Anna. 2008. "First-Person Reports and the Measurement of Happiness." *Philosophical Psychology* 21 (5): 571–83. https://doi.org/10.1080/09515080802412552.

——. 2016. "Is Well-Being Measurable After All?" *Public Health Ethics*, May, phw015. https://doi.org/10.1093/phe/phw015.

——. 2017. *A Philosophy for the Science of Well-Being*. New York, NY: Oxford University Press.

Angner, Erik. 2013. "Is It Possible to Measure Happiness?" *European Journal for Philosophy of Science* 3 (2): 221–40. https://doi.org/10.1007/s13194-013-0065-2.

Ankeny, Rachel, Hasok Chang, Marcel Boumans, and Mieke Boon. 2011. "Introduction: Philosophy of Science in Practice." *European Journal for Philosophy of Science* 1 (3): 303–7. https://doi.org/10.1007/s13194-011-0036-4.

Boumans, Marcel. 2006. "The Difference between Answering a 'Why' Question and Answering a 'How Much' Question." In *Simulation: Pragmatic Construction of Reality*, edited by J. Lenhard, G. Küppers, and T. Shinn, 107–24. Dordrecht: Springer.

——. 2015. *Science Outside the Laboratory: Measurement in Field Science and Economics*. Oxford, New York: Oxford University Press.

Bursten, Julia. 2020. "Lab Report. Lessons from a Multi-Year Collaboration between Nanoscience and Philosophy of Science." In *A Guide to Field Philosophy*, edited by Evelyn Brister and Robert Frodeman, 396. New York: Routledge.

Cartwright, N. Bradburn, and J. Fuller. 2017. "A Theory of Measurement." In *Measurement in Medicine : Philosophical Essays on Assessment and Evaluation.*, edited by L. McClimans. London: Rowman & Littlefield.

Cartwright, Nancy. 1991. "Replicability, Reproducibility, and Robustness: Comments on Harry Collins." *History of Political Economy* 23 (1): 143–55.

Cartwright, and R. Runhardt. 2014. "Measurement." In *Philosophy of Social Science: A New Introduction*, edited by N. Cartwright and E. Montuschi, 265–87. Oxford: Oxford University Press.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chang, Hasok, and Nancy Cartwright. 2008. "Measurement." In *The Routledge Companion to Philosophy of Science*, edited by S. Psillos and M. Curd, 367–75. New York: Routledge.

Collins, H. M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London ; Beverly Hills: Sage Publications.

Desrosières, Alain. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Translated by Camille Naish. Cambridge, MA: Harvard University Press.

Fraassen, Bas C. van. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Frigerio, Aldo, Alessandro Giordani, and Luca Mari. 2010. "Outline of a General Model of Measurement." *Synthese* 175 (2): 123–49. https://doi.org/10.1007/s11229-009-9466-3.

Giordani, Alessandro, and Luca Mari. 2013. "Modeling Measurement: Error and Uncertainty." In *Error and Uncertainty in Scientific Practice*, edited by Marcel Boumans, Giora Hon, and Arthur Petersen, 79–96. Pickering & Chatto.

Hacking, Ian. 1995. "The Looping Effects of Human Kinds." In *Causal Cognition: A Multidisciplinary Debate*, 351–94. Symposia of the Fyssen Foundation. New York, NY, US: Clarendon Press/Oxford University Press.

Hand, David J. 2004. *Measurement Theory and Practice: The World Through Quantification*. Wiley.

JCGM (Joint Committee for Guides in Metrology). 2012. *International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM)*. 3rd ed. Sèvres: JCGM.

Mach, E. 1896. *Principles of the Theory of Heat*. Translated by T. J. McCormack. Dordrecht: E. Reidel.

Mari, L. 2003. "Epistemology of Measurement." *Measurement* 34 (1): 17–30. https://doi.org/10.1016/S0263-2241(03)00016-2.

McClimans, Leah. 2010. "A Theoretical Framework for Patient-Reported Outcome Measures." *Theoretical Medicine and Bioethics* 31 (3): 225–40. https://doi.org/10.1007/s11017-010-9142-0.

McGregor, Thomas, Brock Smith, and Samuel Wills. 2019. "Measuring Inequality." *Oxford Review of Economic Policy* 35 (3): 368–95. https://doi.org/10.1093/oxrep/grz015.

McKay, Andrew. 2002. "Defining and Measuring Inequality." Briefing Paper No 1. UK Department for International Development by the Economists' Resource Centre.

Michell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Ideas in Context. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511490040.

Moscati, Ivan. 2018. *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford Studies in History of Economics. New York: Oxford University Press. https://doi.org/10.1093/oso/9780199372768.001.0001.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. 2. print., and 1. paperback printing. History and Philosophy of Science. Princeton, New Jersey: Princeton University Press.

Ray, Debraj. 1998. *Development Economics*. Princeton, N.J: Princeton University Press.

Reichenbach, Hans. 1927. *The Philosophy of Space and Time*. New-York: Dover, 1958.

Staley, Kent W. 2020. "Securing the Empirical Value of Measurement Results." *The British Journal for the Philosophy of Science* 71 (1): 87–113. https://doi.org/10.1093/bjps/axx036.

Swoyer, Chris. 1987. "The Metaphysics of Measurement." In *Measurement, Realism and Objectivity: Essays on Measurement in the Social and Physical Sciences*, edited by John Forge, 235–90. Australasian Studies in History and Philosophy of Science. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-3919-6_8.

Tal, Eran. 2011. "How Accurate Is the Standard Second?" *Philosophy of Science* 78 (5): 1082–96. https://doi.org/10.1086/662268.

——. 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.

——. 2017. "A Model-Based Epistemology of Measurement." In *Reasoning in Measurement*, edited by Nicola Mößner and Alfred Nordmann, 233–53. London and New York: Routledge.

——. 2019. "Individuating Quantities." *Philosophical Studies* 176 (4): 853–78. https://doi.org/10.1007/s11098-018-1216-2.

——. 2020. "Measurement in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/.

Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge ; New York: Cambridge University Press.