Measurement across the Sciences

Developing a Shared Concept System for Measurement

Luca Mari, Mark Wilson, Andrew Maul

*** SECOND EDITION – FINAL REVISED DRAFT ***

Table of Contents

Preface

All three of us – Luca Mari, Mark Wilson, and Andrew Maul – have dedicated our careers to the topic of measurement, albeit in substantially different areas of application. Luca has a background in physics and engineering, and works in the field of *metrology*, usually defined simply as the scientific study of measurement, but which has until recently almost exclusively entailed the study of *physical* measurement. Mark and Andy have backgrounds in education and psychology, and work in the field of *psychometrics*, usually defined as the scientific study of *psychological* measurement. Given the clear overlap of content, one might expect that the fields of metrology and psychometrics would have a close relationship, and perhaps even that the latter would be a subfield of the former. But this is not the case: indeed, in our experience many professional metrologists and psychometricians are unaware of the very existence of the other field – including the three of us, until our interests happened to lead us each, for different but overlapping reasons, to become concerned with the very possibility of a common foundation of measurement theory and practice.

To put the core issue as briefly as possible: what is it that makes a given process a *measurement*? In particular, can psychological and social properties, such as well-being and reading comprehension ability, truly be *measured*, as is frequently claimed by human scientists and educational testing professionals, and if so, in what way are such processes related to the measurement of physical properties such as length or temperature? Are there shared elements of measurement processes across different domains of application? If so, why have the fields of psychometrics and metrology historically been so disconnected? And if not, are claims about the measurability of psychosocial properties well-justified, or even coherent, given the way measurement is broadly understood in both scientific and lay communities?

To help clarify what we think is at stake with respect to this last question, consider the implications of presenting a given claim or value as being the result of a measurement process. “I have measured your child’s reading comprehension ability” is usually understood as having a different meaning than “I have formed an opinion about your child’s reading comprehension ability”. Relatedly (but not identically, as we will argue), framing claims in numerical terms (e.g., “your child’s reading comprehension ability grew from 80.2 to 91.8 [units] in the past year”) carries different connotations than claims made only in qualitative terms (e.g., “your child’s reading comprehension ability grew substantially in the past year”). Invoking the language of measurement connotes *epistemic authority*: measurement has historically been associated with epistemic virtues such as objectivity, precision, accuracy, and overall trustworthiness, largely as a result of the highly successful history of measurement in the physical sciences and engineering. But, *prima facie*, it is not clear whether measurement processes outside these fields actually deserve to be associated with such authority; a worst-case scenario would be that a given field could be invoking the language of measurement without actually understanding what measurement is, or how its epistemic authority is secured, which in turn could both limit the progress of the field and undermine public trust in measurement in general.

Given this, it would seem that reaching a common understanding of measurement should be treated as an urgent goal. However, as we will explore in this book, measurement is not understood in the same way across different fields of application. Briefly, a weak interpretation of measurement could hold that any process that produces information about a property of an object, according to any rule, counts as measurement: this would have the consequence that, for example, a statement of opinion such as “my well-being has increased during the last year” would qualify. Stronger (i.e., more restrictive) definitions of measurement add further requirements, such as that measurement must

convey relational information of one specific kind, that is, ratios of quantities (with the corollary that a property must be a quantity in order to be measurable); from this perspective, it could be argued that most if not all processes claimed to be measurements in the human sciences are unworthy of the label, because it is not clear that any well-defined quantities or units exist in these fields.

Over approximately the past decade, the three of us have explored these issues together, each motivated by an interest in reaching a common understanding about the nature of measurement. In the course of our conversations, we have each been forced to scrutinize the vocabulary and foundational assumptions of our respective fields. A starting point for our work on this book was our mutual perception that the weak interpretation (as above) was *too weak* – that is, producing information according to any rule is not sufficient for measurement – and the strong interpretation (as above) was *too strong* – at least because we believe the resources of measurement science are also useful in the evaluation of non-ratio properties.

This book presents a summary of the positions we have arrived at as a result of our collaboration. It proposes a concept system about measurement that we believe can be useful to anyone interested in measurement of physical or psychosocial properties. Our proposal, we hope, balances the need for specificity and generality, and as such is indeed stronger than the weak interpretation and weaker than the strong interpretation.

We hope that even if a reader does not agree with everything we propose here, our work will facilitate interdisciplinary communication about measurement (and by extension, science and epistemology in general), and we look forward to the conversations that will follow.

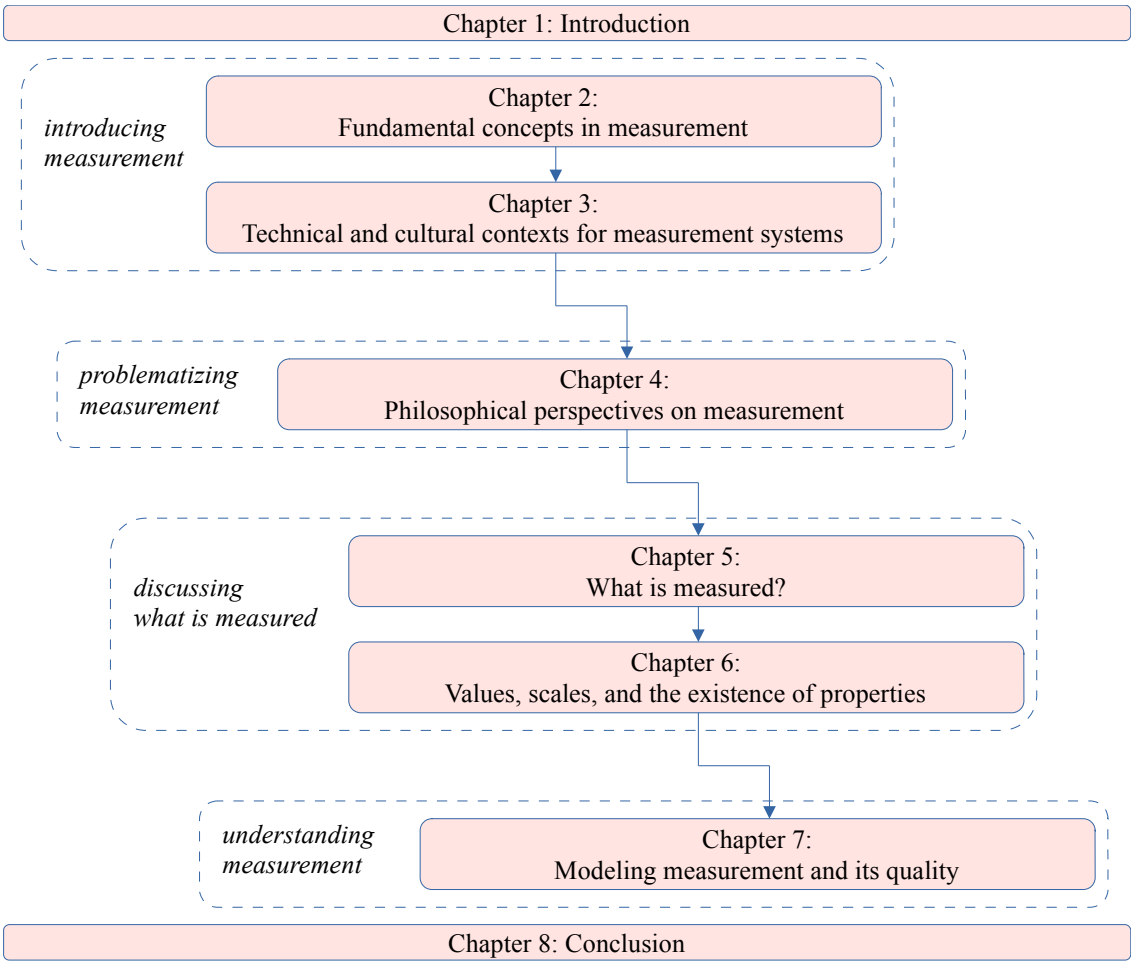
For whom did we write this book?

We see the contents of this book as being of interest to anyone who is concerned with the basic underpinnings of measurement, across (as we indicate in our title) the physical sciences, including biological and geological sciences, and the human sciences, including psychology and sociology, and application areas such as health, education, and management. Moreover, this book should be of interest to anyone who works on developing measuring instruments, especially in fields where the basic foundations may be unclear (such as in the measurement of novel properties in the physical sciences), or where there has been some dispute about basic foundations (such as in the human sciences).

The book could also be used in courses on measurement and/or measurement applications. In designing a course in, say, foundations of measurement, one could use it as the basis for the initial part of the course (with selected extra readings to accompany each chapter), followed by selected readings to focus on particular issues and/or application areas.

The structure of the chapters in this book

The sequence of chapters follows a conceptual path, and therefore we see that the best strategy is to read them from the first to the last. But, at the same time, each chapter was designed to be self-contained and therefore readable independently of the others (including having its own section of references). The arguments that we develop in this book have a simple and basically linear structure:



Chapter 1 sets the stage for the entire work, by introducing and providing some preliminary justification for the very idea that, despite the apparent differences, there could be a way of conceiving measurement across the sciences and fields of application. Chapters 2 and 3 aim to provide a common conceptual and lexical ground for what follows: we believe that their contents are quite basic, so that they could be considered acceptable to all, but that nevertheless someone interested in a simple and structured introduction to these basic aspects of measurement could find something useful there. Chapter 4 proposes our interpretation of the main philosophical conceptions of measurement, and from the strengths and weaknesses we identify, we set up an epistemological background for our endeavor. Given this explicitly philosophical focus, some readers could mainly or even exclusively focus their attention to this chapter, but other readers might skip it instead. Chapters 5 and 6 analyze the role of properties, and more specifically quantities, as objects of measurement. Being a complex subject, about which very diverse positions have been maintained in thousands of years of ontological controversies, we surely do not assume that the realist position presented there will be unanimously endorsed, but do hope that our proposal is acknowledged as consistent and basically fruitful whenever metrological practice has priority over the philosophical presuppositions. Chapter 7 builds upon these pillars and proposes a model of what measurement fundamentally is, highlighting the distinction between direct and indirect (methods of) measurement in reference to the role that models and computation have in the process. We consider it the core of the book. Chapter 8 has been written to offer a view of the path walked through the previous chapters, and leading to a characterization of measurement across the sciences and fields of applications: as such, it could be read as a summary of

the entire book.

As mentioned above, some parts of the book stand alone, and may be useful as such: Chapter 4 on the philosophical conceptions of measurement, Chapters 5 and 6 on properties, and Chapter 8 as an overall summary.

In a complex and dynamic field as measurement science is, no book can claim to have the "last word", and surely this book does not have such an ambition. More humbly, we hope that it will contribute to widening the interest in investigating the foundations of our empirical knowledge, and the critical role that measurement plays in them.

About this second edition

About one year after the publication of the first edition of this book, our editors at Springer kindly proposed that we work on a second edition, to be made openly accessible on the internet. This gave us the opportunity for a careful revision of what we had written. We confirmed the structure of the volume and all its key contents, as in the first edition. Together with the many local changes aimed at improving the text's correctness, clarity, and readability, we expanded the treatment of several subjects, including the pragmatics of measurement (Box 2.2), the distinction between intended properties and effective properties, as also accounted for in terms of definitional uncertainty (Box 2.3), the delicate concept of true value of quantities (Box 6.1), the characterization of what an evaluation scale is (Box 6.2) the main parameters of the behavior of measuring instruments (Section 3.2.1), and the presentation of an exemplary case of application of the Hexagon Framework to psychosocial measurement and instrument development (Section 7.3.5). We also added a new conclusion, "Toward a manifesto for a widespread metrological culture" (Box 8.1), hinting at the social relevance of measurement science, as understood from the perspective of the conceptualization proposed in this book.

Acknowledgments

My research and teaching activities developed along sometimes parallel and sometimes convergent lines, and this was possible thanks to the masters on whose shoulders I stood in these years. This book is an opportunity for paying respect to some of them: Piero Mussio, who for the first time made me feel the charm and the thorns of the scientific research; Mariano Cunietti, who disclosed to me how measurement is a fascinating subject, both scientifically and philosophically; Camillo Bussolati, who decided to involve me, at that time a young PhD student, in a challenging academic endeavor; Ludwik Finkelstein, who widened my then still quite narrow horizons; Sergio Sartori, who accompanied me in discovering the world of metrology; Andrea Taroni, who effectively arranged a context for my work. I devote a grateful thought to the memory of each of them. Coming to the present, my contribution to this book stems from the many conversations that in years I have had with several colleagues and friends. Thank you, Alessandro Giordani, for the long path we walked together in understanding measurement in a philosophical frame. Thank you, Pietro Micheli, for your clever challenges about measurement in organizations. Thank you, Dario Petri and Paolo Carbone, for what we have done together at the threshold between the "theory" and the "practice" of measurement science. Thank you, Jo Goodwin, for what we have done together in the development of Electropedia, the core terminological resource of the International Electrotechnical Commission (IEC). Thanks, present and

past colleagues of Working Group 2 (VIM) of the Joint Committee for Guides in Metrology (JCGM): working for the International Vocabulary of Metrology with you has been an exceptional opportunity for clarifying and sharpening my ideas. Thank you, Chuck Ehrlich: you proved to me that something good may come forth when differences converge. And thank you, Mark and Andy: ten years after our first meeting and hundreds of pages written together, I am even more convinced that interdisciplinarity is the best way to nourish our minds. The internet has embedded us in a wide small-world network: there are many who might see traces of our conversations in this book. It has been a privilege to learn with and from all of you.

Luca Mari

My interest in measurement began when I was teaching Grade 5 at Yarraville West Primary School on the western suburbs of Melbourne, Australia. As a newly-minted teacher, I was surprised that although there were curricula available to help guide teachers about what they should teach, there were no tools available to help me find out what my students knew already. Later, while studying for my PhD with Benjamin D. Wright at the University of Chicago, I came to see that (a) this weakness was one that stretched across almost all of the human sciences, even where much material was available, and (b) that there were indeed sound and practical paths out of there. So, thank you Ben, I appreciate your deep insights into measurement, and on your insistence that the methods of measurement must make sense. Nevertheless, after almost thirty years of working on theory, models and practice in this area, I still felt that there was a gap in the foundations of my approach to measurement, and that is where this volume comes in. Hence, thank you to my two co-authors, Luca and Andy, and to my colleagues, William Fisher and David Torres Irribarra, for persisting with me through many hours (and years) of discussion and effort (and even some disagreements), to lay out this manifesto.

Mark Wilson

The collaboration between the three of us that has taken place over these last years and eventually led to the production of this book has been instrumental in my development as a scholar, and I owe a great debt of gratitude to both Luca Mari and Mark Wilson, the latter of whom was also my PhD advisor. In addition, I am especially grateful to Joshua McGrane, Derek Briggs, and David Torres Irribarra, all of whom are valuable colleagues and invaluable friends. Lastly, I owe everything to Diana Arya, who I met during my PhD studies and has since become not only my wife, but also my closest colleague and intellectual sparring partner. Building both our careers and our lives together has been a privilege beyond measure.

Andrew Maul

Foreword, by Roman Z. Morawski

The idea of measurement standards seems to be as old as our civilization. The documented history of measurements started *ca.* 3000 years ago in Mesopotamia, Egypt and China, where the needs related to the land management and construction of buildings motivated the invention of the first standards of length, area, volume and weight, which next – for centuries – played a very important role in trade, commerce, government, and even religion. The mythical history of measurement is much longer: according to a first-century Romano-Jewish historian Titus Flavius Josephus, it was Biblical Cain who invented weights and measures. After having killed his brother Abel, he went on to commit many other sins, including this terrible innovation "*that changed a world of innocent and noble simplicity, in which people had hitherto lived without such systems, into one forever filled with dishonesty*".[1] Paradoxically, the scientists of the 21st century are more likely to agree with Titus Flavius Josephus than their predecessors because they have fallen prey to bureaucratic systems of research evaluation based on bibliometric indicators, allegedly being measures of research quality…

The Authors of the book *Measurement across the Sciences* have made an attempt to identify a set of basic conditions necessary for measurement, which could be acceptable for most researchers and practitioners active in various areas of measurement application, including both physics and experimental psychology. They have tried, moreover, to find some complementary conditions which are sufficient for correct characterization of measurement. In this way, they have contributed to the endeavors of great methodological significance, *viz.* to the endeavors aimed at drawing a demarcation line between measurements and measurement-like operations. This is a challenge comparable with that of the demarcation problem in the philosophy of science, *i.e.* the problem of criteria for distinguishing science from pseudo-science. Moreover, this seems to be an urgent task in the times when the creative minds of technoscientific milieus are exposed to the influence of simplistic views which are convincingly presented in such books as *How to Measure Anything...*[2] offering five-steps procedures for defining new measurands and new measurement methods for business applications. In light of those guidelines, what was considered a joke 50 years ago may become today a serious business approach to measurement. One of such jokes, most frequently repeated at that time by the students of measurement science, went as follows:

Examiner: "How to measure the height of the university building using a barometer?"

Student: "By offering this barometer to the administrator in exchange for the access to the technical documentation of the building."

Measurements, considered to be the most reliable sources of information, are omnipresent in the life of information society which, by definition, is intensively and extensively involved in the usage, creation, distribution, manipulation and integration of information. The reliable measurement data are indispensable for decision-making processes, especially if the latter are supported by IT tools. The demand for such data appears not only in a research laboratory, but also on a production line and in a hospital. The growing demand for such data may be observed in various institutions of public administration, education and transportation. Unlike in the 19th century, the institutions of business and bureaucratic management are the main driving forces behind the avalanche generation of new

[1] Cited after: Kula, W. (1986). *Measures and men*. Princeton (NJ, USA): Princeton University Press (translated from Polish by R. Szreter), p. 3.

[2] Hubbard, D.W. (2014). *How to measure anything: Finding the value of intangibles in business*. Hoboken (NJ, USA): John Wiley & Sons, Inc. (3rd edition). Hubbard, D.W., & Seiersen, R. (2016). *How to measure anything in cybersecurity risk*. Hoboken (NJ, USA): John Wiley & Sons.

measurands, especially so-called performance indicators, and the corresponding methodologies for their evaluation. Despite the socio-economic damages implied by the reckless application of those indicators for decision-making, despite the common awareness of the so-called Campbell's law[3] and Goodhart's law,[4] their use is not getting less frequent or more prudent. The reasons are obvious:

- they are claimed to be more objective than experts' opinions;
- they may be easily "digested" by the algorithmic procedures supporting the decision-making processes;
- once agreed by the decision-making bodies, they play the role of excuse for pragmatically or morally wrong decisions;
- they effectively replace intellectual qualifications of the decision-makers.

Another driving force of measurement massification is self-tracking biometrics, a growing interest in acquisition of data related to different aspects of our personal life: monitoring of heart condition, mood, air quality, ... This trend towards self-tracking through measurement technology – which appears under the names of body hacking, self-quantifying or lifelogging – is motivated by the promise of a healthier, longer and better life. This promise cannot be fulfilled without rational unification of heterogeneous measurements it relies upon. The book *Measurement across the Sciences* is about such a unification although the idea of self-quantifying does not appear there. Pantometry, i.e., an obsessive desire to measure everything, is another sociological phenomenon – provoked by extensive availability of measurement tools of various quality – which is creating enormous demand for conceptual unification of measurements across various domains of quantities, indicators and measures. Enough to say that the global market of sensors is expected to grow by *ca.* 9% between 2020 and 2025.[5]

The Authors of the book *Measurement across the Sciences* – not succumbing to the temptation of white-black normativeness – provide a very pragmatic answer to a frequently asked question about "bad measurement" by defining it as "*not sufficiently objective and intersubjective according to the given purposes of the measurement*" (Section 7.4.4). It should be noticed, however, that this statement makes sense provided the operation under consideration satisfies at least basic conditions necessary for measurement, identified in the book. Although the title of the book seems to suggest that its contents apply exclusively to measurements in sciences, it actually addresses not only the measurement tools and methodologies dedicated to scientific research, but every instance of measurement which satisfies those basic necessary conditions. One might even risk a hypothesis that the socio-economic impact of the book will be significantly stronger outside of that restricted area – in various domains of engineering, medicine, agriculture, food industry, etc. – where the costs of erroneous decisions implied by ill-defined measurements may be very high.

The book is about philosophical and logical foundations of measurement science. Philosophy is a never-ending discourse on the key assumptions of ontological and epistemic nature, and logic is about systematically deriving conclusions from those assumptions. The Authors have clearly-cut preferences if those assumptions are concerned, but – being aware that they can be justified only *a posteriori* by the distant logical consequences – neither ignore nor oppugn the alternative approaches and views. This is important if the book is to be received not only by philosophers of science – who are inclined

[3] "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (cited after en.wikipedia.org/wiki/Campbell's_law [as of July 20, 2020]).

[4] "When a measure becomes a target, it ceases to be a good measure." (cited after en.wikipedia.org/wiki/Goodhart's_law [as of July 20, 2020]).

[5] Mordor Intelligence: Sensors Market – Growth, Trends, and Forecast (2020–2025), www.mordorintelligence.com/industry-reports/global-sensors-market [as of July 20, 2020].

to invest enormous energy in the unproductive realism-antirealism debates – but also by creative measurement practitioners who are interested in harmonization of various paradigms of measurement, developed and applied in various domains of science and technology. There is such a need, there is such an expectation in the milieus which have to deal with complex systems integrating measurement data representative of the objects, phenomena and events of various nature: physical, chemical, biological, psychological, etc. The book is committed to meeting those needs and expectations. Unlike self-help guides *How to Measure Anything…,* it is not offering ready-for-use solutions, but rather showing the patterns of thinking that could lead to practical solutions of specific classes of problems.

The Authors of the book *Measurement across the Sciences* are affiliated at different institutions and represent complementary fields of expertise related to measurement science: Dr. Luca Mari is Professor of Measurement Science at Università Carlo Cattaneo (Castellanza, Italy), Dr. Mark R. Wilson is Professor of Educational Statistics and Measurement at University of California (Berkeley, USA), and Dr. Andrew Maul is Associate Professor of Education Science and Psychometrics at University of California (Santa Barbara, USA). Before writing this book, they have been involved in a long-term collaboration aimed at making convergent the methodologies of measuring physical and non-physical quantities. Their efforts have had not only scientific but also organizational dimensions: through their efforts measurements in social sciences have been incorporated into the programme of activity of the International Measurement Confederation (IMEKO). Based on the experience of their fruitful collaboration, these highly respected scholars have produced a major work that will be for years to come a central text in measurement science – the text of importance for measurement philosophers, measurement theoreticians and measurement practitioners looking for creative solutions of interdisciplinary problems.

Roman Z. Morawski, Ph.D., D.Sc.
Professor of Measurement Science
Warsaw University of Technology
Poland

Foreword, by Robert J. Mislevy

I work in what is called *educational measurement*: some applications, some methods, some theory. My applications have focused on capabilities people develop in school, work, and recreation (what Herb Simon called "semantically rich domains"), such as standardized tests in science and reading comprehension, and less familiar assessments with studio art portfolios and simulations for troubleshooting computer networks and dental hygienists' procedures. The methods are mainly latent variable models such as item response theory (IRT; more about this later). My theoretical work has been on task design, validity, cognition and assessment, and, our reason for gathering together, measurement. It is from this belvedere, dear reader, that I offer my thoughts on Luca Mari, Mark Wilson, and Andrew Maul's *Measurement across the Sciences: Developing a Shared Concept System for Measurement*.[1] By reflecting on how this system both strengthens and challenges the inquiries of those of us in educational measurement, I hope to share what I find interesting, important, and energizing across any and all disciplines.

Educational Assessment and Educational Measurement

I say that I work in "what is called educational measurement" because most of what most of us do, most of the time, is applications and methods. Millions of assessments every year, producing scores that affect individuals and institutions in ways large and small. This work might better be called *educational assessment*: ways of getting and using information about how people are learning, what they can do, how they think, how they might improve, or how they might fare in educational or occupational settings. Colloquially, the phrase "educational measurement" situates assessment data in a quantitative framework to characterize the evidence that the observations provide for score interpretations and score uses. How one goes about doing this is methods, and we do this a lot. Curiously, far less attention is accorded to more theoretical, more fundamental, questions. Just what kind of *measurements* might those scores be, if indeed they merit the term at all? What relation do purported measures, as numerals and categories, have to attributes of people?

A conjunction of factors, I believe, led to this relative lack of interest. First, assessment itself was familiar. Examinations have been used for more than a millennium, as for selection in civil service in imperial China and matriculation in medieval European universities in Bologna, Paris, and elsewhere. These scores had authority by way of authorities!

Second, the measurement of quantitative properties in the physical sciences had by the turn of the 20th century earned authority the hard way: by theory, evidence, argumentation, instrumentation, and demonstrated coherence within a web of scientific and practical phenomena. *Such a measure holds value not simply because it produces numbers, but because it places a trace of a unique event, observed at a particular time and place in a certain way, into a network of regular relationships among objects and events that holds meaning across times, places, and people.* Being able to calibrate

[1] I have not cited sources rigorously in this more informal preface. I have drawn on Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52(2)*, 161–193; Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge; Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press; Porter, T. M. (2020). *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton University Press; Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation, 23*, 31–48; and others, including the references that appear in Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York/London: Routledge.

local measures to a common metric is a hallmark of physical measurement. There is typically a dialectic between improved theory and improved instrumentation; increasingly efficacious theories of properties and the processes by which they bring about effects under instrumentation (i.e., "opening the black box"). To the shared benefit of science and commerce, national and international institutions such as the *International Bureau of Weights and Measures* standardize units and vocabularies. As two running examples, this book uses length, the canonical case of classical quantitative measurement, and the more nuanced and therefore illuminating case of temperature.

Third, psychologists sought to extend this quantitative measurement frame to the psychological and social realm, or psychosocial sciences in the terminology of this book. Psychophysicists began studying sensory perception and acuity through the lens of measurable human attributes in the mid 1800s, and it is here that philosophical controversies about the nature of measurement in human sciences surfaced. Early psychometricians, such as Charles Spearman and Louis Terman, adapted their methods to data from the emerging standardized tests of educational and psychological constructs such as intelligence, verbal aptitude, and reading comprehension ability (RCA), the running example of this book. My oversimplification: (1) a test was crafted to *educe* a *trait* thought to *exist* as a *property of persons;* (2) persons' interactions with the test situations are coded to produce numbers that are taken to indicate more or less of that property; and (3) the numbers are taken as *measures* of *values* of said *property* for *each person*. This chain of reasoning is replete with terms (italicized above) that could be defined in multiple ways (which have practical consequences), and assumptions for which the strength of support and alternative explanations should be determined.

Serious inquiries into the nature of measurement in psychosocial sciences drew on the substantial foundations established in the physical sciences up to that point, but even so a wide range of views resulted, as seen for example in varied writings of S.S. Stevens, Leon Thurstone, Edward Thorndike, Georg Rasch, Duncan Luce and John Tukey, and Joel Michell. At one pole is that measurement in psychosocial science needs to meet the same axioms as in physical science, producing for individuals' values that correspond to properties of that which is being measured (e.g., a person's RCA), such that the mathematical properties of the values as numbers correspond to demonstrated regularities, within and across individuals, in the real-world interactions they effect (note: "effect" is a stronger word than "affect"). At the other pole is Stevens's pithy definition of measurement as the assignment of numerals to objects and events according to a rule. The task was then to determine just what kind of measurement it might be, in accordance with nominal, ordinal, interval, and ratio properties of numbers. The question of whether a property in question existed was tacitly assumed in the affirmative, often its quantitative nature as well. Between these extremes lies the so-called latent variable measurement models I mentioned earlier. Unsound inferences result when weaker definitions are employed to justify numbers as measurements, but stronger definitions are employed, if tacitly, to justify interpreting them and using them.

Wer fremde Sprachen nicht kennt, weiß nichts von seiner eigenen.

Johann Wolfgang von Goethe

(Those who know no foreign language know nothing of their own.)

Herein lies the genius of this book. Luca, Mark, and Andy step back from the particulars of length, temperature, and increasingly esoteric measures of physical sciences, and purported measures such as corruption, quality of life, and RCA in psychosocial domains. They carefully develop a set of necessary conditions and a set of sufficient conditions they propose to characterize what one might call

proper measurement. They draw on both the successes and earned authority of physical measurement and the points of debate for various ways of extending the notion to psychosocial realms. The concepts and conditions clarify the latter debate and offer a resolution I consider satisfying if not definitive. More interestingly, they offer new insights to the foundations of physical measurement; points of contention in the psychosocial extension bring out distinctions and assumptions that are elided in the development of measurement in the physical sciences, and recorded in standards documents such as the *International Vocabulary of Metrology*.[2]

A compact statement of necessary (but not sufficient conditions) that this book advances is that "Measurement can be considered, preliminarily and in a general sense, to be a process based on empirical interaction with an object and aimed at producing information on a property of that object in the form of values of that property" (Section 2.2). I will not attempt to unpack this statement here, as the authors do so quite thoroughly, but I will note some of the implications they discuss. These necessary conditions encompass classical and derived physical measures. They do not require either values of objects' properties or of empirical observations to satisfy particular mathematical forms. A property characterized by categorical values, for example, could also be shown to be measurable under the framework proposed in this book.

I underscore and will return to two final general implications of their necessity conditions that the authors draw that I consider central to developing a contemporary understanding of educational measurement. The first is existence: *"[T]he critical question is how we* know *that a property exists, and therefore that it meets at least the most basic criterion for measurability"* (Section 6.6.1). The second is that more is needed than a discerned input-output relationship of objects and effects: *"[A]ny purely black box model cannot adequately account for all relevant features of measurement, and thus is not sufficient for the purpose of understanding the quality of measurement results"* (Section 7.1).

Neither will I unpack sufficient conditions for measurability of a property; they too deserve further explanation, which the authors provide quite lucidly. *Objectivity* and *intersubjectivity* are key concepts. Objectivity is the extent to which the conveyed information is about the property under measurement and nothing else (an ideal, to be sure; the authors discuss and illustrate avenues by which it may be examined, critiqued, and quantified in given situations). Intersubjectivity addresses the goal that the conveyed information is interpretable in the same way by different persons in different places and times (similarly explicated and illustrated).

Taking these conditions together, the authors propose taking the necessary and sufficient conditions into account, "*measurement is an empirical and informational process that is designed on purpose, whose input is an empirical property of an object, and that produces explicitly justifiable information in the form of values of that property"* (Section 8.3.4). In particular, *"a property is measurable if and only if there exists a property-related process fulfilling these conditions"* (Section 7.1).

The authors call our attention to the pragmatic nature of measurement as it is actually used, with regard to defining measures, devising instruments, calibrating the values, and justifying their use for given purposes under given conditions (typically less broad in psychosocial domains than physical domains, for reasons we will see in the next section). Objectivity and intersubjectivity are ideals, never perfectly realized. How well can we satisfy them in a particular case, and with what strength of grounding for the nature of a property and the processes by which it produces effects? What is the strength of the arguments, the situated fidelity of the models, and the quality of evidence, *as evaluated within the common framework that this book affords?* Work is required in each case; answers are

[2] Joint Committee for Guides in Metrology (2012). JCGM 200:2012, *International Vocabulary of Metrology: Basic and general concepts and associated terms* (VIM). Sevres: www.bipm.org/en/publications/guides/vim.html.

neither assured nor guaranteed uncontroversial, but we can hold our debates in a shared language and system.

By implication, and as seen historically, advances in theory and instrumentation can strengthen our arguments for a measure and measurements thereof, or instead relegate a project to the dustbin. We may transform our understanding of underlying processes, perhaps opening a black box, to move from a preliminary model to one that is more sophisticated, more explanatory. In one case, emerging theories and laboratory discoveries led early chemists to abandon phlogiston as a property, let alone a measurable one. In another, the quantum revolution revealed how classical measurement of, say, position and velocity jointly, fails for subatomic particles; yet the classical model remains an excellent approximation for measuring classical effects. Advances in technology and learning science have put educational assessment at such a cross roads. Is educational measurement, with RCA as an example, more like phlogiston or a Newtonian measure of velocity?

Opening the Black Box in Educational Measurement

Educational assessment finds itself today in the midst of revolution on multiple fronts, one of which bears most directly on educational measurement. Advances in our understandings of the nature, acquisition, and use of human cognitive capabilities, force reconception of how we might conceive psychosocial attributes that underlie assessment. At the same time, advances in technology, which open the door to new forms of instrumentation, provide opportunities for both improving theories and enrichening applications.

Reading comprehension was a good running example for this book. All readers will have personally experienced comprehension through reading, and most have taken standardized RCA tests consisting of the familiar "read a text and answer questions about it" items. They are intuitively plausible instruments for which performance depends on a presumed property (RCA) that has values, and reading capabilities of individual persons are each characterized by such a value. Test developers learned to produce texts and questions that were fairly predictable in difficulty and fairly reliable in sorting out people who fared better or worse overall. Test-takers' overall scores, usually in the form of sums of correct answers, are considered to indicate, if noisily, their RCAs.

As this book notes, many of the capabilities we address in educational assessment, hence the properties they may be hypothesized to measure, vary over time and place. Tests of RCA obviously vary for assessing test-takers in different languages, but also in different cultures within the same nominal language, and language use itself shows evolution over the decades as to style, popularity of constructions, vocabulary, and topics. What might this mean for "measuring RCA"? Most of the networking certification examinations I worked on with Cisco expired after two years, as new technologies come on line and the meaning of "networking competence" evolves. This is a *social* consideration in determining whether such examinations might assess an existing property; if so, it is one that changes over time, in ways that individuals may or may not. Adding simulation tasks into the certification exams along with multiple choice items redefined what was assessed and what candidates studied. *Cognitive* considerations, I think, are even thornier.

By the early 1930s, Sir Frederic Bartlett began to pry the black box open by proposing schemas, or meaningful patterns of knowledge in the world through which people learn to understand a text, such as folktales, how faces look, or writing a research proposal. Other research showed relationships of item difficulty (as percents-correct in a targeted population) with texts' syntactic complexities and word frequencies (as obtained from pertinent corpora of reading materials). Under these practices, a

reading comprehension test might produce quite serviceable scores for monitoring progress, conducting research, and identifying struggling readers. The relationships of item features to population difficulties served as support for a provisional model of reading comprehension. Now total scores in a collection of similarly-constructed tests will necessarily order persons. But this does not guarantee that there exists a corresponding RCA property of persons, with a range of values, for which each reader possesses a reading comprehension capability that is fully characterized by one of those values. The fact that different publishers' choices of texts, nature of questions, and conditions of performance can reliably sort test-takers differently casts doubt on objectivity and intersubjectivity. Similarly troubling was emerging evidence that some test items could prove differentially difficult for test-takers from different ethnicities, genders, or language backgrounds who performed similarly overall.

A latent variable modeling approach called "item response theory" (IRT), of which the Rasch measurement model illustrated in this book may be considered an instance, provided an analytic framework to make progress on these issues. For familiar correct/incorrect test items, an IRT model gives the probability of a correct response as a function of variables standing for a person's capabilities and variables standing for characteristics such as its difficulty and how well it sorts high and low performers. Under idealized circumstances, these parameters and the form of the model would account for all the systematic variation in a set of responses over some collections of persons and items. The Rasch model in this book is a special case in which, if it were true, the same comparisons in observed performance would hold in probability for people regardless of items, and for items regardless of people. Of course, IRT models are never exactly true in real data such as from science and RCA tests, and for reasons I'll address shortly, we still find systematic variations for item characteristics in different groups of people, and individuals whose response patterns don't accord with the model very well at all.

IRT models have nevertheless proved unquestionably valuable as evidentiary-reasoning frameworks, to improved practice such as enabling adaptive testing for individuals and identifying items which are not functioning as intended. Useful, to be sure, as probability-based machinery to improve practical work. Do IRT models produce measures? Well, IRT models are now being extended to incorporate cognitive theory that connects item features with process-models for what people have to know and do to solve problems. In some cases we can construct items from theory, and predict how they will work. Other more detailed latent-variable models with categorical person variables instead of or in addition to continuous ranges connect even more closely with cognitive findings. The lid of the black box lifts further, with process models beyond those that the trait and behavioral perspectives can provide. We can think of the variables in these models as hypotheses for measurable properties of persons. The statistical relationships between them and cognitively-motivated features of items appears to move us in the direction of *intersubjectivity.*

But an understanding that is emerging across fields as varied as linguistics, sociology, reading research, subject-domain learning, and cognitive, situative, and social psychology opens the box further still. This perspective, which I will call "sociocognitive" for short, provides insights in how people learn, act, and think. It moves us to think more deeply about objectivity, intersubjectivity, and latent variable models in educational measurement.

This sociocognitive perspective[3] is a kind of grand elaboration of Bartlett's schema theory. Every situation we experience, though unique, builds around recurring patterns in language, culture, and subject matter—from elements of grammar and vocabulary, to schemas like Bartlett's, to cultural models, and expectations and behaviors of recurring activities, including taking a standardized test. None of these patterns is a fixed thing; rather they are regularities, with variation, across many individual unique events, across myriad individuals. Similarly, through participating in such events every individual develops resources to recognize the patterns, learn what actions are possible, and create situations around them. Every individual's capabilities are unique, shaped by their history of experience. Comprehension is activating such resources to construct, largely below the level of consciousness, what Walter Kintsch referred to (and this book references) a situation model. Everybody's comprehension of a given text, for example, is personal and unique. A reading passage about a dog walking down a road draws on each reader's history of experiences with dogs, roads, and typical situations in which dogs might walk along roads. Persons from environments in which dogs are friendly pets and those in which dogs are dangerous wild animals will understand the passage differently. This is an example of a reason that questions for a given text can be easy for some people, hard for others, and incomprehensible to still others. An example concerning the choice of questions to ask is that "who, what, where" questions can be answered by intuitive transformations of explicit propositions in a text, while a question about "what is missing?" demands a richer situation model (this question can be the most important aspect of comprehending a real estate contract).

It is a commonplace observation that tasks that are hard for some people can be easy for others, depending on our life experiences. So what does a task variable, like an IRT difficulty variable, actually mean? How can it be a property of an item? Joe Redish, my colleague from the Physics Education Research Group at Maryland, explains IRT as a mean field approximation. Comprehensive explanations of all the unique resources and processes that produced all the responses from all the people in a group to all the items in a collection is overwhelmingly complicated. Yet while people's histories, and therefore their comprehensions, are unique, they are not chaotic. Similarities in peoples' experiences lead to similarities in the resources they develop. To borrow Wittgenstein's term, there are family resemblances across people as to the resources they have developed in relation to certain linguistic, cultural, and substantive patterns, and there are family resemblances across tasks as to the resources people might bring to bear with regard to such patterns. The fit of an IRT model and its person and item variables for a collection of responses to items from people is a probabilistic pattern for the entire ensemble of data. The person variables express, within this framework and data, tendencies of individuals – in the Rasch model, for example, to perform well or poorly. The item variables are grand simplifications of patterns associated with individual items, looking across performances of all people to all items in the ensemble. Together, combining an item's variable with a person variable one person at a time approximates how a person with that value would fare on that item.

Setting aside this IRT story for a moment, a sociocognitive perspective would posit that given each person's constellation of reading comprehension resources, items tend to be harder or easier given, say, the complexity of the item's syntax *as it relates to that individual's history of experience*. In contrast, in IRT this is approximated across all persons by complexity *as it applies in general, ignoring content, context, and individuals' histories*. Similarly, from a sociocognitive perspective, items tend to be harder or easier for a person as its vocabulary relates *to that individual's experience*

[3] More technically, sociocognitive complex adaptive systems (CASs). The special issue of *Language Learning* (Volume 59, Supplement 1) provides a readable overview of CASs, summarizes ways the CAS perspective has revolutionized the field of linguistics, and offers illustrative articles on topics such as the evolution of grammatical structures, children's language learning, and language testing.

with those words and uses. In IRT, word frequencies from corpora take *frequency across the texts as a proxy for word familiarity for each person*. Not really right, but better or worse in some applications, good enough for some purposes in some contexts, and surely a step in the direction of understanding.

Similarity of experiences in a collection of persons that involve both the patterns that are the target of a test and the myriad resources that are also necessary for performance brings us closer to both *objectivity* and *intersubjectivity*: What drives differences in persons' performances is now mainly their resources for the targeted capabilities, and what makes items difficult is similar for everyone involved. As mentioned above, cognitive theory may further provide connections to typical processes and to features in items that predict their variables. Under these idealized circumstances, the person variables of a suitable latent variable model are candidate proxies for values of a property, perhaps further a property that may be argued as measurable. Similar patterns in relationships among items may then arise more widely across persons and tasks, enabling approximate calibrations across such circumstances – a characteristic this book proposes that we require of measurement, to add information beyond the observations at hand. However, we depart more often and more substantially from these necessary conditions as persons become more diverse, as tasks involve more varied knowledge patterns, and as performances become more complex. In a simulation where every action can be logged, for example, we can observe differences among the knowledge and strategies people draw on—differences previously hidden when only simple responses were recorded, indeed differences that further question whether the underlying capabilities can be characterized by different values of the same property.

So, the nature of the capabilities that educational assessments address can vary over time and place, and patterns across peoples' performances can vary in as many ways as there are possible groups of people. Nevertheless, because we develop capabilities in experiences shaped around recurring patterns in knowledge and activity, there are discernable regularities. These regularities reflect recurring patterns in institutions and activities, and we can write books, design instruction, and develop assessments that enable individuals to participate in these activities. In an everyday sense we speak of peoples' RCAs or their competence in computer networking. We can communicate the results of assessment scores, even scores from pools of tasks calibrated through latent-variable models, and use the scores for grading, selection, or certification at large scale. But is this activity *measurement* as we would see it through the lens of the system proposed by this book?

It appears to me that the conditions of existence, objectivity, and intersubjectivity presented in this book must be investigated and evaluated in context. Because of the multifarious constituents of every unique situation (but with "family resemblance" similarities across situations) and the unique personal capabilities of individuals (but with family resemblances that can emerge through experiences with similar constituent patterns) we cannot uncritically expect educational assessments to provide *measures* as per the criteria presented in this book. We can however, examine the degree to which those conditions are approximated, over what ranges of persons and situations, aided by coordinated task development, cognitive theory, and latent-variable modeling machinery. Further, we can identify groups and individuals for whom their patterns in performance are so atypical as to preclude interpretation in the modeling framework. There can be situations for which proceeding *as if* a targeted property exists, is measurable, and is approximated by a given latent-variable model is a satisfactory interpretive frame for most test-takers of interest; it may still be that the performances of some individuals simply cannot be well understood within that framework. (Does the putative property exist for such an individual?)

All of this makes sense to me if I take the system presented in this book, and properties and measures, and variables in latent-variable models as cognitive tools for us, the analysts, to guide our

thinking and our actions. Sometimes we can encounter or construct situations in which the approximation is justified, and it is satisfactory to think and act as if calibrated scores from a latent variable model were measures, even as we remain alert for model misfit and departures from objectivity and intersubjectivity that distort targeted inferences. This book aims to offer an idealized framework for measurement across science. It enables us engineers working in the less-than-ideal real world that we can use to characterize the evidence we provide with regard to its approximation as measures, and thereby improve the quality of our applications.

Conclusion

Some thirty years ago a different cross-disciplinary quest for fundamental concepts across disparate fields transformed my thinking and my career. It was David Schum's decades-long inquiry into whether there might exist "a science of evidence",[4] underlying the reasoning under uncertainty found in varied forms and sources of evidence, concerning disparate inferences and purposes, spanning philosophy, logic, probability, statistics, history, medicine, psychology, and other disciplines. The answer is yes, he concluded. His 1994 text *The evidential foundations of probabilistic reasoning* is the most complete presentation of his findings.[5]

I mentioned my interests in task design, validity, and cognition and assessment. It turns out that all of these can be seen through the lenses of fundamental argumentation structures and reasoning principles, as applied to the evidence, the substance, and the context of the domain of educational assessment. Familiar assessment practices and analytic methods can be re-understood as applications of these more basic structures; likewise the concepts of reliability, validity, comparability, and generalizability. We can use the same framework to address the challenges and opportunities that developments in psychology, technology, and analytics continually present. In each instance, there is serious work to be done: marshalling evidence, constructing arguments, building theory, generating and exploring alternative explanations. It is these foundational structures that embody the deep principles of evidence and inference that underlie *educational assessment.*

It is by further integrating with these evidentiary-reasoning structures the fundamental structures of measurement as this book lays out, and suggests the kinds of evidence and arguments we must marshal with respect to a claim of measurement. We can, through this structure, come to understand whether, and if so in what sense, and to what extent and in what contexts, an activity justifies the interpretations and extrapolations that the very term *educational measurement* implies. This book is that gratifying combination of painstaking philosophy, with beneficial consequences, and a structure that supports the work that must be done in a field like mine that can surely use it.

Robert J. Mislevy
Educational Testing Service
United States

[4] Schum, D. A. (2009). A science of evidence: Contributions from law and probability. *Law, Probability and Risk, 8(3)*, 197–231.

[5] Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Chapter 1.
Introduction

1.1 Why we wrote this book

It would be difficult to overstate the value and importance of measurement in nearly every aspect of society. Every time we purchase or eat food, take prescribed medicine, travel in a vehicle, use a phone or computer, or step inside a building, we place our trust in the results of measurements – and, for the most part, that trust seems well-earned, and as such measurement is commonly associated with precision, accuracy, and overall trustworthiness. Against this backdrop, it seems little wonder that the human sciences[1] have, since their inception, attempted to incorporate measurement into their activities as well. However, despite – or perhaps, to at least some extent, because of – the ubiquity of measurement-related concepts and discourse, there remains a remarkable lack of shared understanding of these concepts across (and often within) different fields, perhaps most visibly reflected in the vast array of proposed definitions of measurement itself. In addition to hampering communication across different disciplinary fields regarding shared methodological principles, such a lack of common understanding hints at the possibility that the same terms – "measurement", "measurement result", measurement model", etc. – are used with very different and possibly even incompatible meanings in different disciplines, with potentially disastrous results.

Of course, measurement is not a natural entity, pre-existing and waiting to be discovered; rather, measurement is designed and performed on purpose. Hence, in attempting to define or characterize measurement, we inevitably must attend to domain-related conventions and customs in the contexts in which measurement has been developed and established. Given the aforementioned lack of common understanding of measurement across the scientific and technical literatures, one might conclude that there is an irreducible multiplicity of measurement-related concepts and terms; from this perspective, an endeavor aimed at exploring a possible shared understanding of measurement across the sciences would seem to be pointless.

Obviously this is not our position. We believe, instead, that a transdisciplinary understanding of the nature of measurement is a valuable target, for both theoretical and practical reasons. As previously noted, measurement is commonly acknowledged to be a (or even *the*) basic process of acquiring and formally expressing information on empirical entities, which suggests the usefulness of a shared understanding of basic and general concepts related to measurement (hence not only <measurement> itself, but also <measurand>, <measurement result>, <uncertainty>, <accuracy>, etc.[2]). Information is routinely acquired and reported by means of values on properties as diverse as reading comprehension ability, well-being, the quality of industrial products, the complexity of software systems, user satisfaction with social services, and political attitudes. But should these cases all be understood as instances of measurement? Stated alternatively, are all such cases worthy of the trust commonly afforded to measurement, and if so, what do they share in common that makes them so worthy? Or, at

[1] We use the term "human sciences" to refer to all scientific disciplines and activities concerned with the human mind and behavior, including not only psychology, but also sociology, anthropology, and disciplines of research concerned with particular activities such as education, health and medicine, economics, and organizations. Thus, the term is interpreted analogously with the term "physical sciences," which refers not only to physics but also other disciplines concerned with physical phenomena, such as chemistry, biology, geology, and astronomy.

[2] We will carefully distinguish here *objects*, *concepts* for understanding objects, and *terms* for designating objects, and some notation will support us in this: thus, for example, measurement (no delimiters), as an object, is understood via a concept <measurement> (angular brackets) and designated in English by the term "measurement" (double quotes). See Box 2.1 for a presentation of our terminological assumptions.

least in some cases, are such examples better understood as instances of something other than measurement – perhaps something less trustworthy, such as statements of opinion or subjective evaluation? To restate the issue as succinctly as possible: what justifies the perceived *epistemic authority* of measurement?

We think any attempt to answer such questions will require acknowledgment that measurement is not something that can be isolated from scientific and technical knowledge more generally. Characterizations of measurement found in different fields and different historical periods relate in important ways to more general issues in science, philosophy, and society, such as the nature of properties and the objects that bear them, the relationship between experimentation and modeling, and the relationships between data, information, and knowledge – and indeed, the very possibility of (true) knowledge. In particular, on this last point, a question of increasing relevance to our data-saturated world is: given the increasing trends of interest in "big data" and "datafication", under what conditions does data actually provide information on the measurand? In the radically new context of widespread availability of large, sometimes huge, amounts of data in which we are now living, it is plausible that measurement science will maintain a role in our society only by attaining a broadly shared fundamental basis, instead of dissolving in a myriad of technical sub-disciplines.

1.1.1 Is measurement necessarily physical?

Thomas S. Kuhn (1961: p. 161) once observed that

> at the University of Chicago, the facade of the Social Science Research Building bears Lord Kelvin's famous dictum: "when you cannot measure, your knowledge is of a meagre and unsatisfactory kind." Would that statement be there if it had been written not by a physicist, but by a sociologist, political scientist, or economist? Or again, would terms like "meter reading" and "yardstick" recur so frequently in contemporary discussions of epistemology and scientific method were it not for the prestige of modern physical science and the fact that measurement so obviously bulks large in its research?

It is hard to dispute that, for most, the paragon of measurement is *physical* measurement. For some, this might even be the end of the conversation: measurement is necessarily of physical quantities, and thus anything called "measurement" in the human sciences is either ultimately of something physical, or is at best a metaphorical application of the concept of measurement to something that is in fact not measurement. And indeed, there is some historical weight to this argument: for much of the history of human civilization, measurement was associated with a relatively small number of spatiotemporal properties, such as length, mass, and time, and more recently force, temperature, and electric charge. As scientific understanding of the physical world has advanced, these properties have become increasingly understood as mutually interdependent, via physical laws (such as Newton's second law of motion, which posits that force is the product of mass and acceleration); when values are attributed to physical properties, such laws can be used for inferential purposes by operating mathematically on the available values by means of the relevant laws. Reasoning about the physical world in this way proved so successful that it was the common ground upon which new branches of physics were created in the 18th and 19th centuries, in particular thermodynamics and electromagnetism, the development of each of which involved the discovery of their own sets of properties and laws connecting them. And, of course, such scientific advances led to technological changes, which in turn triggered further scientific advances, as well as changes in society at large.

This positive feedback loop would not have been possible without effective tools for obtaining information about the relevant properties. This is, of course, the role played by measurement; as

Norman Campbell effectively summarized, "the object of measurement is to enable the powerful weapon of mathematical analysis to be applied to the subject matter of science" (1920: p. 267).[3] Measurement is thus a critical component of the scientific paradigm of the physical sciences, and has been integral to its success.

Again, given this, it is perhaps unsurprising that other scientific fields and areas of human activity have increasingly incorporated measurement-related concepts and terms into their own activities. In part, this seems to be based on a widespread acceptance of the premise contained in Lord Kelvin's credo – that without measurement, knowledge is at best "meagre and unsatisfactory" – with the further implication that, as put for example by psychologist James McKeen Cattell, "psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement" (1890: p. 373; for more extended discussions see, e.g., Michell, 1999, and Briggs, 2021).

In the late 19th and early 20th centuries, scholars working in the field of psychophysics, such as Gustav Theodor Fechner and Stanley Smith Stevens, investigated relationships between physical stimuli and their associated human responses (sensation and perception), on the premise that by establishing quantitative relationships (e.g., through what is now known as Fechner's Law; Fechner, 1860) between known physical quantities such as weight and human sensations, the latter could be measured as well.[4] Separately, scholars interested in individual differences, such as Francis Ysidro Edgeworth (1888; 1892) and Charles Spearman (1904; 1907), applied statistical methods and logic originally developed in the context of astronomy (in particular related to the Gaussian distribution) to the study of human beings, largely in the context of scores on educational and psychological tests. Starting from the observation that some sets of test items, such as arithmetic questions, seemed to give more internally consistent results than others, they posited that an individual's "observed scores" (O) on tests could be decomposed into "true scores" (T) and "errors" (E), i.e. $O = T + E$, giving rise to a field that would later come to be known as *psychometrics*.

Perhaps unsurprisingly, these early attempts at measuring psychosocial properties were met with skepticism by some members of the broader scientific community. As one important committee concluded, referring in particular to psychophysics, "to insist on calling these other processes measurement adds nothing to their actual significance but merely debases the coinage of verbal intercourse" (Ferguson et al., 1940: p. 345). Even within the human sciences, many were (and still are) skeptical of psychosocial measurement, often based on concerns such as that "not everything that can be counted counts, and not everything that counts can be counted", as put by the sociologist William Bruce Cameron (1963: p. 13).

In part, skepticism about psychosocial measurement may have been related to the very fact that, as previously mentioned, available examples of measurement pertained exclusively to physical properties, which might have given the impression that only physical properties are measurable. Interestingly, even many prominent proponents of psychosocial measurement seem to have accepted this position; for example, Jum C. Nunnally and Ira H. Bernstein, in the 3rd edition of their influential textbook *Psychometric Theory* noted that "it is more defensible to make no claims for the objective reality of a construct name such as 'anxiety' and simply use the construct name as a convenient label for a particular set of observable variables. The name is 'valid' only to the extent that it accurately describes the kinds of observables being studied to others. [...] The words that scientists use to denote

[3] This position seems to be broadly accepted in the human sciences as well; for example, as put by Warren Torgerson, "measurement enables the tool of mathematics to be applied to science" (1958: p. 1).

[4] Work in the psychophysical tradition helped establish a number of relationships between physical phenomena and sensations (e.g., Cattell & Fullerton, 1892; Thurstone, 1927) that are even today still foundational for fields such as audiology.

constructs (e.g., ‘anxiety’ and ‘intelligence’) have no real counterparts in the world of observables; they are only heuristic devices for exploring observables. Whereas, for example, the scientist might find it more comfortable to speak of anxiety than of [item] set *A*, only set *A* and its relations objectively exist, research results relate only set *A*, and, in the final analysis, only relations within members of set *A* and between set *A* and members of other sets can be unquestionably documented.” (Nunnally & Bernstein, 1994) This use of a term such as “heuristic devices” again seems to imply that claims about measurement in the human sciences are best understood as being metaphorical rather than literal; alternatively, one might conclude the term “measurement” simply has irreducibly different meanings in the physical sciences and the human sciences, which was indeed the conclusion of some human scientists like Stanley Smith Stevens (see, e.g., McGrane, 2015)).[5]

But, to us, such conclusions seem unsatisfactory: again, measurement is regarded as integral to science and society on the basis of its *epistemic authority*, and so the question remains of what, exactly, justifies claims to such authority. As Kuhn asked: “what [is] the source of [the] special efficacy” of measurement? (1961: p. 162). Stated alternatively, what are the necessary and sufficient elements of trustworthy measurement processes, independent of domain or area of application? We hope, in this book, to address exactly this question: *what could be a common foundation of measurement across the sciences?*

1.2 Some familiar and not-so familiar contexts for measurement

In the sections below, we introduce two examples of the sorts of measurement that we had in mind when writing this book. Each will appear later at several points in the text, along with other examples when they are more pertinent. In particular, we recognize that many of our readers might not have experience with measurement across both the physical sciences and the human sciences, and hence the accounts are each designed to be quite basic, starting from a very low expectation of expertise in their respective topic areas. These basic accounts will be expanded, deepened and updated at appropriate places in the text. We have also included a third section, where we give an illustration of how the typical format of measurement in the human sciences, in terms of sets of items, can be seen as structurally analogous to measurement approaches in the physical sciences.

1.2.1 A brief introduction to temperature and its measurement

While discussing the features and the problems of measurement systems in this book, we mention some examples of physical properties, including the well-known cases of length and mass. In particular, in Chap. 6 the hypothesis that length is an additive quantity is exploited in the construction that, starting from lengths of rods, leads to units of length and, hence, values of length. But a bit more is developed for the example of temperature, which is used in Chap. 6 for showing how values may be obtained for a non-additive quantity and also in Chap. 7 where we introduce a model of direct measurement.

From the perspective of our conceptual analysis of measurement, temperature has some very interesting features. It is, first, a property of critical importance: “Temperature has a profound

[5] A more nuanced variation of this conclusion is that there are different kinds of measurement that share common properties; this was the conclusion reached in particular by Ludwik Finkelstein, who argued for a distinction between “strongly defined measurement” and “weakly defined measurement”, where the former “follows the paradigm of the physical sciences [and] is based on: (i) precisely defined empirical operations, (ii) mapping on the real number line on which an operation of addition is defined, (iii) well-formed theories for broad domains of knowledge”, and the latter is “measurement that [...] lacks some, or all, of the above distinctive characteristics of strong measurement” (Finkelstein, 2003: p. 42).

influence upon living organisms. Animal life is normally feasible only within a narrow range of body temperatures, with the extremes extending from about 0–5 °C (32–41 °F) to about 40–45 °C (104–113 °F).".[6] It is a property that we perceive with our senses and that we understand qualitatively, in relative terms of warmer and colder, but the quality of our perception system is quite low, in particular due to its limited selectivity (what we actually perceive is the so-called *apparent temperature*, caused by the combined effects of air temperature, relative humidity, and wind speed) and range (our thermoception loses all discriminatory power for temperatures outside the narrow range mentioned above). Given its practical importance, it is not surprising that the history of the understanding and the measurement of temperature is rich, with several significant stages (see, e.g., Chang, 2007, and Sherry, 2011), from the starting point of our physiology, which allows us to consider temperature only as a (partially) ordinal property based on the relation *warmer than*, to the introduction of instruments which make differences of temperature observable by transducing temperature to the height of a liquid via the effect of thermal expansion. Such instruments were based on the hypothesis of a *causal relation* between temperature and volume: *ceteris paribus*, if the temperature of the liquid increases then its volume increases (and then also its height increases, thanks to the ingenious configuration of the instrument). In other words, the problem of the low sensitivity to temperature of the human senses was solved not by looking for some sort of "temperature amplifier", but by gaining and then exploiting knowledge about the effects of temperature on a second property, which is in some sense more directly observable, by means of instruments that were designed as *nomological machines* (Cartwright, 1999).[7]

The discovery that a transduction effect can be effectively modeled as a monotonic relation is sufficient for building instruments that only detect changes of the relevant property, as is, for example, the case for a *thermoscope,* which is a device able to reveal a change of temperature by somehow showing a change of volume. However, the *ceteris paribus* condition is critical for a quantitative characterization of the transduction effect, and therefore equipping a thermoscope with a scale and thus making it a *thermometer*, given the dependency of the transduced height on the context – air pressure in particular – and the instrument's features, including the kind of liquid used and the material of which the tube is made (typically some sort of glass). It was only on the basis of such a condition that fixed points were discovered, so that, e.g., *ceteris paribus*, water boils always at the same temperature. This was a fundamental enabler of the establishment of scales of temperature, which were initially created without a strong theoretical understanding of temperature and its relation to thermal expansion, and instead were mainly based on models of data, typically with the assumption of linearity of values between the fixed points (Bringmann & Eronen, 2015). The compatibility of the results produced by different instruments was hard to achieve, and in consequence so was the construction of a socially agreed thermometric scale, Celsius and Fahrenheit being only the two remnants of a larger set of once-proposed scales. But this multiplicity of instruments, able to produce at least partially compatible results, also helped advance our knowledge of temperature: the observed transduction effects implemented in different instruments share a common cause, which is also the same physical property that we perceive and describe in terms of warmer or colder. This standpoint was further supported by the discovery of other temperature-related transduction effects, independent of thermal expansion, for example the thermoelectric effect, such that differences of temperature are transduced to differences of electric potential (i.e., voltage). The hypothesis of the existence of

[6] www.britannica.com/science/thermoreception.

[7] The same effect of thermal expansion was also exploited as the basis of instruments whose nomological structure is more complex, like the so-called Galileo thermometers, in which changes of volume produced by changes of temperature in turn produce changes in density (i.e., mass divided by volume), which are made observable as changes in buoyancy conditions, e.g., of bodies that become denser than the fluid in which they are immersed, and thus eventually do not cease to float if the temperature of the fluid increases.

temperature, as the cause of multiple, independent but correlated effects, was thus strongly corroborated.

Temperature has some other interesting features for our conceptual metrological perspective. It is an *intensive* property, i.e., "one that is independent of the quantity of matter being considered",[8] so that the temperature of a thermally homogeneous body does not change by removing a part of the body, and nevertheless it has a fundamental connection with several additive / extensive properties, and in particular heat energy, which spontaneously flows from bodies at a higher temperature to bodies at a lower temperature. Moreover, the temperature of a gas is equivalent to the average kinetic energy of its molecules, where thus a property at the macroscopic level (temperature) is explained in terms of a property at the microscopic level (molecular kinetic energy).

Finally, the measurement of temperature and its development are also interesting with respect to scale types. While historically temperature was considered to be only an ordinal property, the scientific and technological advances resulting from the adoption of the experimental method led to thermometric scales (including the previously-mentioned Celsius and Fahrenheit scales), which are interval scales, because of the lack of knowledge of a "natural zero", common to all scales. (Compare to the case of length and mass: even though many scales of length and mass were introduced, each corresponding to a different unit, all scales of length share the same zero-length and all scales of mass share the same zero-mass.) Accordingly, ratios of values of (thermometric) temperature are still not meaningful – in the sense that if the temperatures of two bodies are, e.g., 20 °C and 40 °C, then the conclusion that the latter is twice as warm as the former is mistaken, as one can easily check by converting the two values to °F – but units of temperature are nevertheless well-defined, and allow us to compare invariantly the ratios of *differences* of values. A second scientific development created the conditions for the final step: thermodynamics implies the existence of a minimum, or absolute zero, of temperature, at –273.15 °C, which led to the Kelvin scale, and which is thus a ratio scale, although, of course, still non-additive, as further discussed in Sect. 6.3.6.

1.2.2 A brief introduction to reading comprehension ability and its measurement

An important example of measurement in the human science domain is that of a student's reading comprehension ability (RCA). The relevance of reading comprehension ability to the modern world can hardly be exaggerated; indeed, you, the reader, would not have gotten this far without your own reading comprehension! It is obvious that accurate measurement of RCA is of crucial importance in education, but it is equally so in many other social domains, such as in the writing of guidebooks, the formulation of tests such as driving tests, and in the communication of public health warnings.

A basic scenario for the measurement of reading comprehension ability might involve the following.

(a) A reader reads a textual passage, and is then asked one or more questions about how well they understand the contents of the passage. One of the first such tests was developed by Frederick Kelly (1916), and an example question from that test (see Fig. 1.1) will serve as an illustration of this typical format. The questions were chosen by Kelly to be likely to generate incontrovertibly correct or incorrect responses. Such questions and their accompanying rules for interpretation of responses are commonly called *items* in this field.

I have red, green and yellow papers in my hand.

[8] www.britannica.com/science/temperature.

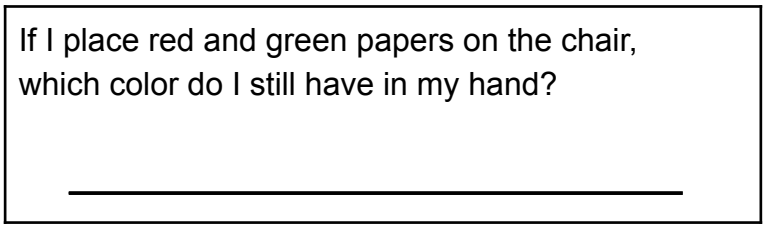


Fig. 1.1 An item from Kelly's reading test

(b) The reader responds to each item by writing a response or selecting an option from a predetermined set of options. Thus, the reader's RCA is transduced to the responses to the items.

(c) A rater judges the correctness of each item response (this may be carried out automatically, especially in the case of multiple choice items).

(d) An initial indication of a reader's RCA is then given by the pattern of correct and incorrect responses for the set of items, which might be summarized in terms of the number (or percentage) of test items that the reader answered correctly, typically called the "sum-score". The sum-score is then an indication at the macro level of the reader's comprehension of the reading passage, whereas each individual item response is an indication at the micro level of the reader's comprehension of the question asked in the item.

(e) This indication is, of course, limited in its interpretation to just the specific set of items administered. A variety of methods are available to allow interpretation beyond that specific test to generate a value on an instrument-independent RCA scale (some of which are presented below).

In a typical educational context, once the measurement is completed, a teacher would use interpretational curriculum materials keyed to the RCA scale value to assign the reader to some specific reading instruction activities designed to be appropriate for her level of RCA.

If the process just described has been successful, the basic input of the process is the student's RCA, and the basic output is the estimated value of the student's reading comprehension ability. As in the case of temperature, other inputs are usually present that could contribute to the output, such as distracting noises, the mood of the student, peculiarities of the text passages and/or the questions, and even specific background characteristics of the reader, etc.

A traditional method to generate an RCA scale is the norm-referenced approach, which relates the RCA values to the sum-score distribution for a chosen reference population of readers. In this method, a representative sample of individual readers from a specified population (e.g., five-year-olds in a given country) take a test, and this generates a sample of results in the form of the local values ("sum-scores") on the test. Then some statistics are computed on the sum-scores, such as the mean and the standard deviation (or the median and the interquartile range, or the percentiles), and the public reference properties are taken to be the RCAs of readers at those values. For example, if the mean and standard deviation were the chosen reference points, then a scale could be set by mapping the mean to, say, 500, and the standard deviation to, say, 100: thus, following this scale formulation, a value of 600 RCA units would be for a reader located at one standard deviation (100 RCA units)[9] above the mean (500 RCA units). This is thus an ordinal scale, but it is often treated as an interval level scale in psychosocial measurement. An alternative approach to the norm-referenced approach is where the RCA scale values are related to specific reference reading comprehension criteria the is known as the criterion-referenced approach, and an example of that will be discussed in Sect. 7.3.5.

[9] More typically, these would be called "points": i.e., "100 points above the mean".

1.2.3 An initial view of psychosocial measurement from a physical science perspective

Some may find it difficult to relate the above account of an RCA test to the traditional idea of a physical instrument such as a thermometer. Seeing the analogies can be not so obvious – in particular, it can be hard to conceive how observations of how well readers respond to RCA items can be compared to how temperature is reflected in a thermometer – as these just do not seem like similar events!

This subsection (itself based on Mari & Wilson, 2014) is intended as a stepping-stone between these two world views of measurement. It starts with a standard physical measurement context, specifically the measurement of temperature using an alcohol thermometer, and shows how this can be adapted to a situation analogous to that for RCA.

With the aim of measuring a temperature Θ, a thermometer can be exploited as an indicating measuring instrument, and specifically as a sensor, which is supposed to behave according to a transduction function (sometimes also called "observation function": this is what Fechner called a "measurement formula", Boumans, 2007: p. 234) assumed to be linear in the relevant range:

$$x = \Theta / k \qquad (1)$$

i.e., the measurement principle is that an object a of temperature $\Theta[a]$ put in interaction with a thermometer of sensitivity k^{-1} generates an expansion of the substance (e.g., a gas or a liquid) in the thermometer bulb and therefore an elongation $x = \Theta[a] / k$ of the substance in the tube.[10] (We will omit measurement units from now on, but of course we take the kelvin, K, as the unit of the temperature Θ, the metre, m, as the unit of the elongation x, and K m^{-1} as the unit of the constant k.) Once the instrument has been calibrated, and therefore a value for k is obtained, the measurement is performed by applying the measurand $\Theta[a]$ to the thermometer, getting a value for the indication x and finally exploiting the inverted version of the law, $\Theta[a] = k\,x$ for calculating a value for the measurand.

This relationship (which is linear, due to constant sensitivity) is illustrated in Fig. 1.2 by the dotted line.

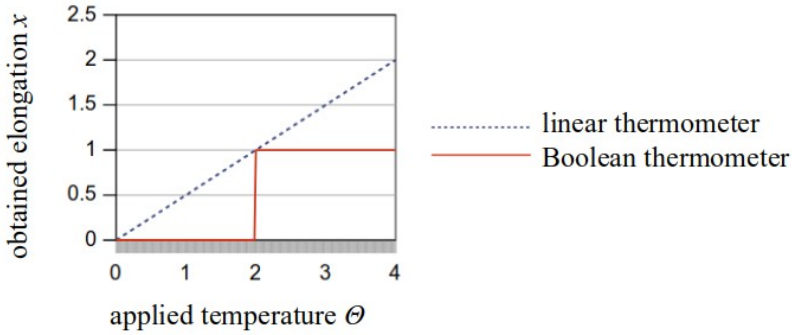


Fig. 1.2 The relationship between Θ and x (the transduction function) for thermometers as given in eq. (1) and eq. (2) (scaled values)

Suppose now that, instead of a thermometer whose behavior is described by eq. (1), a modified thermometer is available, again characterized by a constant k, operating according to the following transduction function:

[10] The notation $\Theta[a]$ is introduced and explained in Footnote 15 of Chapter 2.

$$\begin{cases} \text{if } \dfrac{\Theta[a]}{k} < 1\text{, then the alcohol stays in its rest position, } x = 0 \\ \text{if } \dfrac{\Theta[a]}{k} \geq 1\text{, then the alcohol elongates to a fixed position, } x = 1 \end{cases} \quad (2)$$

Let us call such a transducer a "Boolean thermometer", whereas "linear thermometer" will be the term for any transducer behaving according to eq. (1). (The principle of transduction for a Boolean thermometer is not important here: we might suppose, for example, that the substance enters the tube only when it reaches its boiling temperature – as such, it could be interpreted as a calibrated thermoscope.)

While the behavior of a linear thermometer is mathematically modeled as a continuous, linear function in the relevant range, eq. (2) defines a function whose range is discrete, and in fact binary. A second major difference between eq. (1) and eq. (2) is related to the dimension of the parameter k: while in the case of linear thermometers, dim Θ / k = dim x = L, and therefore dim k = Θ L^{-1}, eq. (2) assumes that dim Θ / k = 1 (i.e., is a quantity with unit one – sometimes the term "dimensionless quantity" is used in this case), so that dim k = dim Θ for Boolean thermometers. The fact that in this case the parameter k is dimensionally homogeneous to a temperature has the important consequence that it can be interpreted as a "threshold temperature", such that the substance elongates in the tube only if the applied temperature is greater than the threshold. This interpretation is crucial for what follows, as it allows the comparison of the involved quantities not only through ratios (Θ / k > 1) but also through differences ($\Theta - k > 0$) and orderings ($\Theta > k$), and therefore makes it possible to place values of the measurand and the parameter of the measuring instrument on the same scale.

Calibrating such a Boolean thermometer requires one to apply increasing temperatures whose values are known and registering the value Θ' of the temperature that makes the substance elongate, so that $k = \Theta'$. If we then apply a temperature Θ to this calibrated Boolean thermometer, and we obtain the indication value x = 1, then the only conclusion that can be drawn in this case is that $\Theta / k \geq 1$, and therefore that $\Theta \geq k$. Thus, this Boolean thermometer has taken the underlying algebraically rich scale of the quantity subject to measurement (the temperature Θ) and rendered it as an ordinal quantity: that is, it has operated as a pass/fail classifier. This, then, is the link to the correct/incorrect nature of the RCA items, as described in the previous section. Of course, the imperfection of this instrument is clear – it does no more than divide up temperatures into two categories, above k and below (or equal to) k. And this is, of course, also the reason why RCA tests always consist of multiple RCA items: so that the RCA scale will be able to distinguish more categories.

Then, to accomplish this using Boolean thermometers, suppose that an array of M calibrated Boolean thermometers is available, each of them with a different constant k_i, and sequenced so that $k_i < k_{i+1}$ (this sequencing is an immediate by-product of the calibration described in the previous paragraph). Then, given an object to be measured a, the measurement procedure would be to apply the measurand $\Theta[a]$ to the Boolean thermometers in sequence until the j-th thermometer is identified such that:

- $\Theta[a]$ generates an elongation in all thermometers i, $i < j$, i.e., the indication value $x_i = 1$ is obtained, so that $\Theta[a] \geq k_i$;
- $\Theta[a]$ does not generate an elongation in the j-th thermometer, i.e., the indication value $x_j = 0$ is obtained, so that $\Theta[a] < k_j$

Hence if j = 1, i.e., no thermometers elongate, $\Theta[a] < k_1$, and if $j = M+1$, i.e., all thermometers elongate, $\Theta[a] \geq k_M$.

In the simplest case of a sequence of $M = 2$ Boolean thermometers, with constants k_1 and k_2, $k_1 < k_2$, three cases can then arise:

(a) $x_1 = 0$, i.e., the applied temperature does not elongate any Boolean thermometer: $\Theta[a] < k_1$;

(b) $x_1 = 1$ and $x_2 = 0$, i.e., the applied temperature elongates the first Boolean thermometer but not the second one: $k_1 \leq \Theta[a] < k_2$;

(c) $x_2 = 1$, i.e., the applied temperature elongates both the Boolean thermometers: $\Theta[a] \geq k_2$.

Clearly, this procedure can be extended to any number of Boolean thermometers that were pragmatically usable in a given context. And that is exactly the formal foundation for one classical approach to measurement in the human sciences, called Guttman Scaling (1944). Under this approach, RCA "Guttman" items are seen as being related to the underlying scale as is the Boolean thermometer in eq. (2), and a sequence of successively harder Guttman items are generated, so that they specify an ordinal scale of readers.

This illustrates what one can do if one already has an algebraically rich measurand such as temperature. The real situation in the case of RCA is, of course, that this is not readily available, so that one must, in some sense, reverse the logic that was worked through here, to proceed from the Guttman items back to the underlying scale. The problem is actually a little bit more complicated than that, as the drawback to this formulation is that RCA and other human sciences items only seldom behave so exactly as given in eq. (2), but rather they function in a less reliable way, so that an element of probability must be introduced in order to better model the situation. In fact, one way to do so is illustrated in Fig. 1.3, where the indication is given in terms of a probability. How this can be done is described in Sect. 7.3.7.

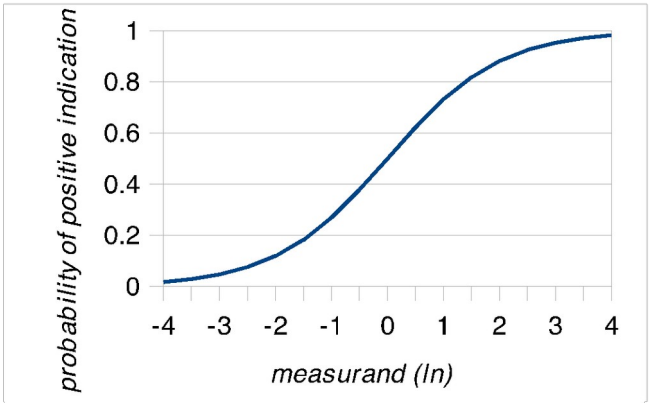


Fig. 1.3 A sketch of a transduction relationship between an RCA measurand (in log metric) and the probability of observing a correct response

1.3 The path we will travel in this book

As we say above, in this book we are seeking a conceptualization of <measurement> that can encompass evaluation of both physical and psychosocial properties, and of both quantitative and non-quantitative properties. In doing so, we require that this conceptualization is specific enough to account for the acknowledged epistemic authority of measurement, which is a critical part of its importance.

We start our story proper in Chap. 2, where we seek to identify a basic set of conditions necessary for measurement, which we hypothesize to be acceptable for many, if not all, researchers and

practitioners across a wide range of fields of application of measurement. Chapter 2 concludes with a statement that summarizes those conditions:

> measurement is an empirical and informational process, designed on purpose, whose input is an empirical property of an object and that produces information in the form of values of that property

In Chap. 3 we will add to this position three key additional points. First, we stipulate that *measurement results should include information about the quality of the reported values*, though we acknowledge that sometimes this is neglected in non-scientific situations. Formerly, this has been considered in reference to measurement errors, but, in contemporary measurement, it is more usually characterized in terms of uncertainty and validity, in physical and psychosocial measurement respectively.

Second, as inherited from the Euclidean tradition, when we report measured values, we are providing a relational form of information – the ratio of the measured property to the chosen unit, in quantitative cases. To do so requires that there is broad social availability of a metrological system that disseminates the reference properties by the usual means of measurement standards connected through traceability chains. This means that *measurement requires calibration*.

Third, in our view, and despite our previous point, we see that there has been an overemphasis on the relevance of the Euclidean tradition to measurement science. In particular, this tradition refers to a concept that is only loosely related to the above-mentioned empirical and informational process of measurement: the mathematical concept <measure>, i.e., a numerical ratio of entities. Hence, our conclusion is that *the contention that measurement applies only to quantitative properties cannot be justified by kowtowing to the Euclidean tradition*.

At this point in the book, we begin our own explorations beyond these basic positions, and address the question: *given these necessary conditions, what complementary conditions are sufficient to characterize measurement?*

As a background to answering that question, we review, in Chap. 4, three broad perspectives on measurement – realism, operationalism, and representationalism – and discuss, in the context of each of them, the epistemic status of measurement and the conditions of its proper use. We present the main findings of this discussion in a simple two-by-two mapping,[11] and the whole discussion leads us to the conclusion that an essential characterization of measurement is as an empirically structured model of the process, rather than some set of mathematical constraints on the inputs or the outputs of the process. This, coupled with an acknowledgment of the inevitable role of models in the measurement process, can be summarized as a *model-dependent realism about measurement*.

Next, in Chap. 5, we take up the very target of measurement, i.e., properties. We analyze properties from both ontological and epistemological perspectives, and identify a core issue in terms of the meaning of the Basic Evaluation Equation (BEE),

> property of a given object = value of a property

which displays the basic components of any measurement result, and which must also be complemented with some information about uncertainty. From our model-dependent realist standpoint, we interpret the BEE relation as the (simple though controversial) claim of an *actual referential equality*: the BEE conveys information on the measurand because the measurand and the measured value remain conceptually distinct entities, though they identify the same individual property. Our position that measurement is an empirical process forces us to conclude that properties cannot be conceptual entities, and hence we must investigate the very existence of properties. An additional

[11] As in Fig. 4.6, whose axes highlight whether measurement has been characterized as being dependent on empirical and/or mathematical constraints, respectively.

complexity of the subject of the existence of properties is that <property> is a cluster concept, including four sub-concepts:

- <property of an object> (e.g., the mass of a given object and the reading comprehension ability of a given individual);
- <value of a property> (e.g., 1.234 kg and 1.23 logits on a specific RCA scale);
- <individual property> (e.g., a given mass and a given RCA);
- <general property> (e.g., mass and RCA).

In our realist perspective, individual properties exist as universals, but the interpretation of the BEE as a referential equality is compatible with other positions: hence, the continued progress in our exploration is not thwarted by possible disagreements over the actual nature of the entities exemplified by one or more of the sub-concepts of <property>.

In Chap. 6 we discuss three fundamental issues for measurement science. The first issue concerns the nature of values of properties, though we start by discussing the values of quantities. We provide a step-by-step construction to show that values are not symbols for the representation of properties, but that *values are individual properties, identified as elements of a scale*. Taking this perspective, we can see that the difference between values of quantitative and non-quantitative properties is a matter of the structure of the scale to which they belong. The second issue is then about the structure of scales and the related conditions of invariance, so that scale types provide a classification for property evaluations and then properties themselves. Our analysis finds no unique condition for separating quantitative and non-quantitative properties, and this finding reinforces the *distinction between being quantitative and being measurable*. The third issue in this chapter concerns general properties. Our basic assumption here is that an empirical process can interact only with an empirically existing entity, and that this applies both to the objects that bear the properties and the properties of the objects. Thus, a distinction needs to be maintained between empirical properties and the mathematical variables that may be used as mathematical models of properties. Regarding the conditions of the existence of general properties and the possible role of measurement in the definition of general properties, *the hypothesis of existence of an empirical property can be corroborated by the observation of effects causally attributed to the property*.

In Chap. 7 we reach the high point of our story where we propose a general model of the measurement process, one consistent with the ontological and epistemological commitments developed in the chapters before. Again, we start with the distinction between empirical and informational processes, and recall that measurement is neither a purely empirical nor a purely informational process. We broadly distinguish between direct and indirect methods of measurement as a fundamental classification of measurement methods related to the complementary roles of these empirical and informational components: indirect measurements necessarily include at least one direct measurement. In consequence, we give *a structural characterization of direct measurement as the actual foundation of measurement science*. This structural characterization we call the *Hexagon Framework*, and we exemplify it for both physical and psychosocial properties. We also use the Framework to highlight the importance of evaluating the quality of the information produced by a measurement, but now frame this in terms of the high-level, complementary conditions of object-relatedness ("objectivity") and subject-independence ("intersubjectivity"). Finally, the Framework provides a sufficient condition for measurability*: a property is measurable if it is the input of at least one process that has been successfully structured according to the Framework*.

Thus, in the conclusion of our story in Chap. 8, we revisit the arguments and discussions of the earlier chapters. We come back to address our initial question: following the necessary conditions we

discuss in Chaps. 2 and 3, and the conclusions we reach in the subsequent chapters, what sufficient conditions, complementary to the necessary conditions, do we propose for characterizing measurement across the sciences?

References

Boumans, M. (2007). Invariance and calibration. In M. Boumans (Ed.), *Measurement in economics: A handbook* (pp. 231–248). London: Academic Press.

Briggs D. C. (2021). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. New York: Routledge.

Bringmann, L. F., & Eronen, M. I. (2015). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*. doi.org/10.1177/0959354315617253

Cameron, W. B. (1963). *Informal Sociology: A casual introduction to sociological thinking*. New York: Random House.

Campbell, N. R. (1920). *Physics – The elements*. Cambridge: Cambridge University Press.

Cartwright, N. (1999). *The dappled world – A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373–380.

Cattell, J. M., & Fullerton, G. S. (1892). The psychophysics of movement. *Mind, 1(3)*, 447–452.

Chang, H. (2007). *Inventing temperature – Measurement and scientific progress*. Oxford: Oxford University Press.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51,* 599–635.

Edgeworth, F. Y. (1892). Correlated averages. *Philosophical Magazine, 5th Series, 34*, 190–204.

Fechner, G. T. (1860/1966). *Elements of Psychophysics*. New York: Holt, Rinehart & Winston.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Report of the British Association for the Advancement of Science, 2*, 331–349.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement, 34*, 39–48.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Kelly, E. J. (1916). The Kansas silent reading tests. *Journal of Educational Psychology, 7*, 63–80.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52*(2), 161–193.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement, 51*, 315–327.

McGrane J. A. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology, 6*, 431.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Nunnally, J. M., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A, 42*, 509–524.

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

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161–169.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–386.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Chapter 2.
Fundamental concepts in measurement

This chapter aims to present the general context of a measurement system and basic concepts of measurement and related terms. The presentation develops according to a step-by-step, top-down strategy, which progressively characterizes measurement as (i) an empirical process, (ii) designed on purpose, (iii) whose input is a property of an object, and (iv) that produces information in the form of values of that property. These are proposed as necessary but not sufficient conditions for a process to be identified as a measurement: as such the contents of this chapter should be uncontroversial to be read and accepted by most, if not all, researchers and practitioners.

2.1 Introduction

Measurement is present in nearly every aspect of modern society, from handicraft to large-scale scientific research and from trade and commerce to complex technology. Perhaps because of this omnipresence, it sometimes seems that measurement is just taken for granted. However, the basic concepts and terms associated with measurement are not universally shared or agreed-upon; in different contexts, different terms are sometimes used for the same concept, or the same term is used for different concepts. Just as an example, in its common usage the term "measure" can denote the entity to be measured (as in "length is a geometric measure"), the process of measuring (as in "this measure has been hard to perform"), the result of the process (as in "this measure is sufficient for making a decision"). Moreover, there are many stereotypes and presuppositions related to measurement, particularly regarding its relations with quantification (Mari et al., 2017). To help dispose of these stereotypes and presuppositions, we plan in this book to lay out a basic set of foundations for measurement that we believe will be useful across a wide range of sciences and applications. The fact that measurement is so deeply embedded in the infrastructure of our society seems sufficient to suggest that a mutual understanding of its basic foundations would be very useful, even if it is quite challenging to achieve.

Any framework that encompasses *measurement across the sciences* requires a unified and consistent terminology – i.e., a coordinated set of concepts and the terms to designate them. This is a complex endeavor given the conceptual and lexical multiplicity around measurement, in which one frequently encounters both homonyms (one term, several meanings: a serious problem, due to possible misunderstandings) and synonyms (one concept, several terms: a less serious problem, just a matter of confusing redundancy).

The importance of developing a basic terminology of measurement has already been broadly acknowledged, and several international organizations have been cooperating in pursuit of this goal for some decades in the Joint Committee for Guides in Metrology[1] (JCGM), one of whose outcomes is the

[1] The current member organizations of JCGM are: the two inter-governmental organizations concerned with metrology: the Bureau International des Poids et Mesures (BIPM) and the Organisation Internationale de Métrologie Légale (OIML); the two principal international standardization organizations: the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC); three international unions: the International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP), and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC); one international accreditation organization: the International Laboratory Accreditation Cooperation (ILAC) (JCGM, 2009).

International Vocabulary of Metrology (the so-called VIM, from its French title, *Vocabulaire International de Métrologie*; JCGM, 2012), the other being the *Guide to the expression of uncertainty in measurement* (the so-called GUM; JCGM, 2008). The JCGM documents aim "primarily at harmonizing worldwide current metrological practices and disseminating scientific and technological knowledge. They constitute recommendations that member organizations are strongly encouraged to implement." (JCGM, 2009: A.1.2). Of course, such recommendations should apply more to institutional tasks[2] than to scientific research, which is expected to perform free exploration not bounded by prescriptions or proscriptions. Nevertheless, the VIM is a well established and widely used document, and therefore, to the extent possible, we adopt it as a basic reference here. In particular, we adhere to, but go beyond, one of its assumptions: "In this Vocabulary, it is taken for granted that there is no fundamental difference in the basic principles of measurement in physics, chemistry, laboratory medicine, biology, or engineering. Furthermore, an attempt has been made to meet conceptual needs of measurement in fields such as biochemistry, food science, forensic science, and molecular biology." (JCGM, 2012: Introduction).

With this book we aim at paving a way for the addition of the human sciences – including psychology, sociology, and economics, as well as fields of application such as education, health, and management – to this list. Hence, in what follows we present our attempt to further expand the scope of the fundamental principles of measurement, so as to include both physical and non-physical measurement[3] in a single, consistent concept system, grounded on a sound terminology, as introduced in Box 2.1.

This requires us to depart from the VIM, and actually from a long-standing tradition in measurement, for two structural reasons:

[2] An example is legal metrology, the "practice and process of applying statutory and regulatory structure and enforcement to metrology" (OIML, 2103: 1.01), that is required to produce standardized documents such as the European Union's Directive of Measuring Instruments (EU, 2014).

[3] The distinction between what is physical and what is not is complex, and touches the fundamental problem of reductionism (can chemistry be considered a part of physics? biology? etc.), which is a key subject of philosophy of science, but which can safely remain in the background in a discourse on measurement science. We avoid a systematic use of the term "non-physical" here (and not only for political correctness: characterizing something in negative terms does not necessarily convey a clear meaning), and use instead the adjectives "human science" and "psychosocial", in a broad sense, as attributed to a science, a measurement, a property, etc., to emphasize that that entity is not effectively defined in purely physical terms. Of course, some non-physical measurement may not be human (e.g., behavior of dogs), and some human measurement may be entirely physical (e.g., height), but we are not concerned with such cases here. A discussion on the compatibility of reductionism with the acknowledgment of the possibility of multiple layers of description is in Philip Warren Anderson's (1972) paper *More is different*, whose main thesis is twofold. On the one hand, "the reductionist hypothesis [is that] the workings of our minds and bodies, and of all the animate or inanimate matter [...], are assumed to be controlled by the same set of fundamental laws". On the other hand, "the reductionist hypothesis does not by any means imply a "constructionist" one: the ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe" (p. 393). In other terms, a principled reductionism can be maintained together with the acknowledgment that effective descriptions of parts of the world are given in reference to non-reduced parts. Daniel Dennett has proposed a high-level interpretation of this subject in terms of *physical* vs. *design* vs. *intentional* stance (Dennett, 1987).

- we discuss measurement in terms of *properties*,[4] whereas the VIM assumes that only *quantities* are measurable;
- we include *psychosocial* properties in the system, whereas the VIM assumes that only *physical* quantities are measurable.

Both of these differences imply a generalization, whose justification is a core task of this book. A summary characterization of the fundamental concepts introduced in this chapter and used through the whole book is in Appendix "A basic concept system of measurement".

We distinguish the contents of this and the following chapter from the rest of the book by noting that we believe the points and positions adopted here should be generally and broadly acceptable to measurement experts of many types. Thus, these two chapters are intended to set up the broad background for the remaining chapters. This chapter aims at presenting the general context of a *measurement system* and introducing some basic concepts of measurement and the related terms.

Box 2.1 – Entities of the world, concepts, and terms: a primer on terminology and concept systems

We describe the world by means of language, with the aim of producing shareable knowledge. This requires that terms in descriptions are understood in the same way by all subjects involved in the communication, that is, there is an *intersubjective* agreement as to their meaning. It is then "useful for our purposes to distinguish between *concepts* [...] and the corresponding *terms*, the verbal or symbolic expressions that stand for those concepts" (Hempel, 1966: p. 275) (a term is not constrained to be a single word: for example, "measurement uncertainty" is one term), where concepts are "units of knowledge" (ISO, 2000: 3.2.1), that, in order to be communicated, processed, and stored, require a linguistic form. Hence a relation between language, knowledge, and the world is implied:

- knowledge is about the world: for example, the concept <measurement> is intended to be about actual measurement processes;
- knowledge is managed by means of linguistic expressions: the concept <measurement> is expressed by the term "measurement" in English and "mesurage" in French;
- if knowledge is properly established and shared, then both the English "measurement" and the French "mesurage" designate actual measurement processes.

The relations between language, knowledge, and the world, and more specifically between terms, concepts, and entities in the world, are effectively depicted in the so-called "triangle of reference", or "*semiotic triangle*" (Ogden & Richards, 1923), as in Fig. 2.1:

[4] The concept is so fundamental that, not surprisingly, together with "property" several other terms are used to designate it, with meanings more or less analogous, like "attribute", "feature", "characteristic", "quality", "observable", "parameter", etc. The differences in standpoints about properties are not only lexical: some approaches to measurement avoid discussion of properties, by dealing only with empirical objects, represented by means of informational entities (usually but not necessarily numbers) through procedures. Whether properties do exist in the world, or are just conceptual tools we adopt to organize our knowledge of empirical objects, is a core topic for a fundamental ontology (see, e.g., Orilia & Swoyer, 2020) and deeply affects any measurement-related concept system (for example, do we measure objects or properties of objects?). In this book we maintain the usual position that what is measured are properties of objects, like the mass of solid bodies and the reading comprehension ability of individuals, and therefore that properties of objects exist, and are therefore not concepts. This position is developed further and defended in Chap. 5. See also the summary in Table 2.1 below.

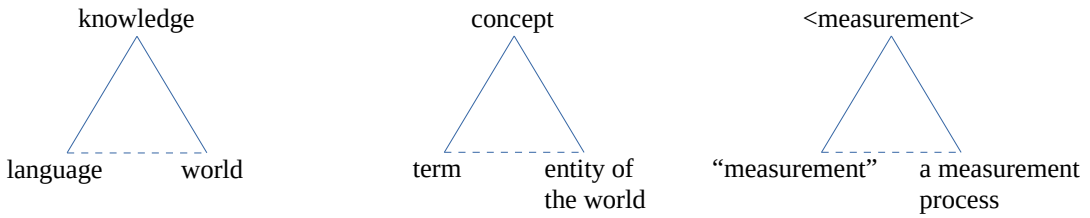


Fig. 2.1 The semiotic triangle, in the generic case (left) and the specific case (middle), with an example (right). Adapted from Ogden & Richards (1923: p. 11)

This model is so fundamental that, perhaps unsurprisingly, the related terminology is not standardized. For example, instead of “term, concept, entity of the world”, Charles Ogden and Ivor Richards (1923) use “word, thought, thing”, and Mario Bunge (1974: p. XI) uses “symbol, construct, fact”, whereas ISO terminological standards (ISO, 2000; ISO, 2009) use “designation, concept, object”.

A term usually stands for the *entity* it designates, but sometimes we use terms to refer to *concepts* or even to *terms* themselves. For example, on the subject of measurement one could write that it is a process aimed at producing quantitative information: here the reference is to the entity-measurement, a process performed by means of suitable instruments etc. One could also write that <measurement> is intended differently in physical and social sciences: here the reference is to the concept-measurement, a critical unit of knowledge in metrology. Finally, it might be even the case that one wants to emphasize that “measurement” is an English noun, the plural form of which is “measurements”: here the reference is to the term-measurement, a lexical entity.

The *sense* of a term is the concept it designates, and the *referent* of a term is the entity it refers to. An analogous distinction is put between the intension and the extension of a concept: “The set of characteristics that come together to form the concept is called the *intension* of the concept. The set of objects conceptualized as a concept is known as the *extension* of the concept.” (ISO, 2009: 5.4.3).

In order to maintain a clear distinction between terms referring to entities of the world, to concepts, and to terms, we will henceforth adapt – as done in the example above – the notational convention of ISO standards (e.g., ISO, 2009). A term, and more generally a linguistic expression, referring to:

- itself, i.e., a term, is delimited by double quotes “ ”;
- a concept, i.e., the meaning of the term, is delimited by angle brackets < >;
- an entity of the world, i.e., the referent of the term, is not delimited,

where the lack of delimiters around terms for entities of the world is in accordance to the economic principle that terms for entities of the world should not be delimited because in everyday writing we usually refer to entities of the world, not to concepts or terms.

Hence, the sentence

<measurement> is expressed in English by “measurement” and is about measurement

is correct and means that the concept <measurement> is expressed in English by the term “measurement” and is about measurement as an entity of the world. (It may appear that there is exception in expressions such as

the concept of measurement

but this implicitly stands for

the concept of the entity of the world measurement

and thus the term is not delimited.)

Throughout the book we adopt this notation (the change from ISO standards is in the delimiters for concepts: while ISO standards recommend single quotes, e.g., the concept ‘measurement’, we use angle brackets, which are more clearly distinguished from double quotes), with the only exception of italic or bold terms, which are not delimited. Furthermore, quotations are also delimited by double quotes “as in this example”, and quotations inside quotations are delimited by single quotes.

2.2 The abstract structure of measurement

Measurement can be considered, preliminarily and in a general sense, to be *a process based on empirical interaction with an object and aimed at producing information on a property of that object in the form of values of that property*.[5, 6, 7]

This characterization is generic: it provides conditions that are deemed to be necessary but not also sufficient. In fact, not every such process is a measurement, and several other definitions have been proposed that add empirical constraints on the process and/or formal constraints on the entities considered to be measurable (see the critical review in Chap. 4 and Mari, 2013). Nevertheless, some key features of measurement are already inherent in this characterization: we introduce them here with a step-by-step, top-down strategy, by imposing more and more constraints and therefore progressively specifying the scope of measurement, in terms of the following four black-box conditions.

Measurement:

- is an empirical process (Sect. 2.2.1),
- designed on purpose (Sect. 2.2.2),
- whose input is a property of an object (Sect. 2.2.3), and
- that produces information in the form of values of that property (Sect. 2.2.4).

Since these conditions are necessary, each of them restricts the set of candidate processes to be identified as measurements – for example, the first condition implies that anything that is not an empirical process cannot be a measurement. A more specific characterization of measurement requires that these necessary conditions be complemented with sufficient conditions, as introduced and discussed in Chap. 7.

[5] Compare this to the more specific definition of <measurement> in the VIM: “process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity” (JCGM, 2012: 2.1). Our use of the indefinite article “a” rather than the definite article “the” (“*a* process…”, not “*the* process…”) emphasizes that in this chapter we introduce *necessary* but not *sufficient* conditions to characterize measurement.

[6] In this characterization the term “object” is used in a broad sense and refers to what the VIM calls a “phenomenon, body, or substance” (for example in JCGM, 2012: 1.1) but also to an individual, an organization, a process, etc. There is no claim that an object, in this sense, is a single, unitary entity, and instead it may be a system, or even a conglomerate, of parts. Furthermore, “object” is ambiguous, given that in semiotics an object is meant to be an object of the discourse and therefore properties are, in this sense, objects. Without any further specific ontological commitment, we consider here an object to be *anything that bears properties*; in Chaps. 5 and 6 we develop a more analytical consideration of properties.

[7] The term “value” is ambiguous. In particular, a phrase such as “to be of little value to somebody” shows that it can be used, as an uncountable, for <the quality of being useful or important>, as mentioned in the Oxford English Dictionary. In measurement science, and in this book, “value” has a meaning analogous to what in mathematics is <element of the range of a function> (so that, for example, 1 is the value of the function $\cos(x)$ when applied to the argument $x = 0$), thus devoid of any ethical or axiological components. However, an ambiguity remains: if X is a variable ranging over a set $\{x_i\}$, each x_i is said to be a value of X; hence by modeling a property, such as mass or shape, as a variable, each of its instances, such as a given mass and a given shape, would be considered to be a value of that variable. As we discuss in the following pages, and further in Chap. 6, in the tradition of measurement science the term “value of a property / quantity” is reserved for the instances of properties / quantities that are identified as elements of a classification, which in the case of quantities is induced by the appropriate composition of a quantity chosen as the unit.

Given the ambiguities mentioned above about the basic concepts of measurement, it may be useful to summarize some of our background assumptions and terminological choices, as listed in Table 2.1.

Table 2.1 Some background assumptions

Objects (e.g., physical bodies, phenomena, events, human beings, organizations, systems, ...) have *properties*. For example, a solid body has a mass, a shape, ...; a human being has an age, a reading comprehension ability, ...; a reading event has a duration, a comprehension outcome, ...
Some objects are *comparable* with respect to some of their properties (or, equivalently and more succinctly, properties of objects are comparable). For example, two human beings are comparable by their masses and by their reading comprehension abilities, but the mass of an individual cannot be compared with the reading comprehension ability of another individual. Properties that are comparable are said to be *of the same kind*, being instances of the same *general property*: comparable properties are *individual properties*. For example, two given masses are individual masses, instances of the same general property mass.
Some properties have a quantitative structure,[8] i.e., they are *quantities*. For example, mass is a quantity, whereas shape is a property but not a quantity. Properties that are not quantities may admit of ordering among the objects (e.g., a preference among options) or at least a classification among the objects (e.g., a classification according to shape of bodies): these are called *ordinal properties* and *nominal properties* respectively.
Properties of objects that are empirical entities may be modeled as variables taking *property values*. For example, "this body is a cube" is a shorthand for "the shape of this body is cubical", stating that the body has a shape, that is then a property of an object, which is cubical, hence a shape, i.e., a value of shape. If properties are quantities, their values, i.e., *quantity values*, are customarily (possibly non-integer) multiples of the unit[9] for the quantity. For example, 1.2345 kg is a value of mass, where the kilogram is the unit of mass.
Measurements are processes aimed at producing information on properties of objects in the form of property values, and therefore in the form of quantity values if the property is a quantity.

Some consequences of these assumptions are listed in Table 2.2.

[8] What characterizes a property as a quantity is a delicate subject: Chap. 6 is devoted to discussing it as well as other topics.

[9] The usual term – "measurement unit" or "unit of measurement", as in the VIM (JCGM, 2012: 1.9) – misleadingly conveys the idea that values of quantities only come from measurement, a plainly false position (one can guess that a given rod is longer than 0.123 m, with no measurements implied in the production of this result). Hence, we use instead the term "quantity unit", or "unit of quantity", which are also easier to reconcile with linguistic customs, such as "unit of length": indeed, length is a quantity, not a measurement.

Table 2.2 Some consequences of the assumptions listed in Table 2.1

All quantities of objects are properties, but there are properties of objects that are not quantities.
Properties of objects and values of properties (and therefore in particular quantities of objects and values of quantities) are distinct: properties of objects are identified through empirical means (typically by somehow referring to objects that bear them), while values of properties are identified through formal / mathematical means (typically as multiples of a unit in the case of quantities, and more generally as related to elements of a scale).
As we reserve the term "measurement" to refer to a process, we call "measurand" the property (or the quantity) intended to be measured and, where needed, distinguish an *individual* measurand, the property of a given object that is intended to be measured, and a *general* measurand, the general property that is intended to be measured. The term "measurement result" refers to the information entity (usually one or more property values, but sometimes more complex results like probability distributions over the set of values) produced by the process. We avoid the ambiguous term "measure" as a noun.

These assumptions and terminological choices are discussed and justified in what follows.[10]

2.2.1 Measurement as an empirical process

The first black-box condition of measurement is that it is an *empirical process* that operates on inputs to produce outputs, as depicted in Fig. 2.2.

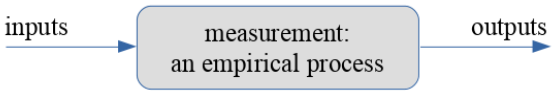


Fig. 2.2 The abstract structure of measurement (first version)

The condition that measurement is empirical aims primarily at demarcating it from informational processes, such as computation and logical inference, and thought experiments: computing a mathematical function produces a value but is not a measurement; conceiving of an experiment that produces values of quantities is not measurement. This is not as trivial as one might suppose. Particularly in the context of geometry, the distinction between measurement and computation is in fact sometimes confused, and the computation of the length of segments or of the areas of surfaces is typically called a measurement (Lockhart, 2012). Furthermore, with the widespread use of numerical methods based on computers, the idea of purely computational, and therefore non-empirical, experiments, for example as performed through simulation, is now common. Hence, characterizing measurement as an experimental process – as the VIM does (JCGM, 2012: 2.1) – does not seem to be

[10] Even though these terminological choices are very preliminary, they are not void of content. In particular, while we maintain that properties of objects may be empirical entities, *modeled* as variables, sometimes properties and variables are not formally distinguished, "for economy of notation", as in the case of the GUM (JCGM, 2008: 4.1.1, Note 1), or perhaps because the difference between empirical entities and their mathematical counterparts is neglected. A telling example is found in the following sentence: "By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or *values*, from a specified domain." (Pearl, 2009: p. 8), which also includes the term "measurement" plausibly in the sense of <measurand>. While this sentence may make sense in Pearl's terms, it is clearly not consistent with the definitions above.

correct (also considering that an experiment could be a *thought* experiment, and thinking is not sufficient for measuring...): measurement must be an *empirical* process.

The condition that measurement is empirical imposes a constraint on the entities that can be considered as its inputs, and therefore as candidates for measurement: it must be possible to interact with them in an empirical way,[11] again a key condition to differentiate between measurement and computation. In some cases this distinction is subtle. For example, is software an entity with which we interact empirically? No, if a software program is considered as consisting only of code, and therefore ultimately a sequence of zeros and ones; yes, if the program is considered as a process executed by a given hardware system in a specific context. Accordingly, while for example the number of lines of code of a software program is something which is computed rather than measured, the effectiveness of the user interface of the same program, once it is executed, depends on several empirical conditions, and therefore it is an empirical property that might be the object of measurement, not computation.[12]

Of course, this condition is not sufficient to characterize measurement: there are plenty of input-output empirical processes that are not measurements. An example is combustion, in which oxygen is chemically combined with other substances (inputs) and heat and light are produced (outputs). Clearly, combustion is not measurement. However, we propose that all examples of measurements can be uncontroversially characterized as input-output empirical processes.

In the simple case of the measurement of human body temperature by means of a mercury thermometer, the basic input is the temperature to be measured and the basic output is the information obtained on this temperature, in the form of a value such as 36.8 °C. More generally, other inputs could contribute to the output, including the temperature and the atmospheric pressure of the environment; an intermediate output is the final position of the upper surface of the mercury in the glass tube of the thermometer as the result of the interaction between the individual and the thermometer, and a final output of the measurement would be the temperature in degrees Celsius, whereas the final outcome of the whole process could be the assessment whether the considered human being is sick or in good health.[13]

[11] This is generally a two-way interaction, and hence the inputs of a measurement may be affected by their being measured. In reference to the traditional distinction between observation and experiment, not all experiments are measurements, and some measurements are only specific kinds of observations, whenever the measured property is unaffected by its being measured. A paradigmatic example is the case of the measurement of the spectral characteristics of the electromagnetic radiation emitted by stars: a star does not change its state because of this process. However, an intervention might be required on the object under measurement before the measurement and in preparation for it, an operation sometimes called “signal conditioning”, with the aim of making the property measurable as expected. For example, electrical resistance measurement typically assumes that a potential difference has been applied, thus in fact changing the state of the system. This justifies the idea that usually measurement is a kind of experiment. Other perspectives on measurement are possible. According to a slightly different construal (which still connects measurement and quantification, a relation about which we provide some critical comments in the following), “there are three modes of generating data: by observation, measurement, and experiment. Observation, whether direct or with the help of instruments and theories, is deliberate and controlled perception, and it is the basic mode of data generation. [...] Measurement [...] may be characterized as quantitative observation, or the observation of quantitative properties. Experiment [is] the observation (and possibly measurement) of changes under our partial control.” (Bunge, 1983: p. 91).

[12] We do not *define* here the distinction between measurement and computation, though we aim instead to provide a pragmatic characterization. In the words of Percy Bridgman, “There are certain human activities which apparently have perfect sharpness. The realm of mathematics and of logic is such a realm, par excellence. Here we have yes-no sharpness. But this yes-no sharpness is found only in the realm of things *we say*, as distinguished from the realm of things *we do*. Nothing that happens in the laboratory corresponds to the statement that a given point is either on a given line or it is not.” (1959: p. 226, emphasis added). According to Bridgman’s metaphor, measurement is something “we do”, and computation is something “we say”. This helps point up the paradigmatic contrast between the exactitude of computation and the uncertainty of the empirical activities of measurement.

[13] If the final output of the measurement process were the decision about the person’s health state rather than the temperature value, then this would be an example of a non-quantitative evaluation, whose values might be, e.g., healthy, rather sick, and seriously sick. We return to a discussion of the conditions under which such an evaluation might be considered a measurement in Sect. 6.5.

An example in the human science domain is the measurement of a student's reading comprehension ability (RCA), which illustrates some of the complexities that can underlie the concept of empirical input-output process. A typical scenario for the measurement of RCA was given in Sect. 1.2.2, and might result in an estimate of the student's RCA in the form of a value on a RCA scale, and a teacher might use interpretational materials to assign the student to a category that involves some specific reading instruction activities. The basic input here is the student's comprehension of the raw text, and the basic output is the estimated value of the student's RCA. As in the case of temperature, other inputs are usually present that could contribute to the output, such as distracting noises, the mood of the student, etc.; an intermediate output is the set of student responses to the comprehension questions, and a final outcome could be seen as the category of instructional activities assigned by the teacher.

2.2.2 Measurement as a designed process

The second black-box condition of measurement is that it is a process *designed on purpose*, rather than a transformation that spontaneously happens. This has the consequence that measurement is performed according to specifications, called a *measurement procedure*, as depicted in Fig. 2.3.

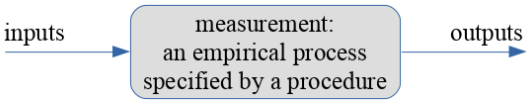


Fig. 2.3 The abstract structure of measurement (second version)

This is an example of the terminological care we must take here: while "process" and "procedure" are sometimes used as synonyms, we follow the VIM and distinguish the transformation that occurs in the measurement from the *normative description* of the transformation. That is, a measurement is a process that is performed according to a specification, and a measurement procedure is a specification that is implemented in a process.[14] This assumes that measurement is inherently pragmatic: it is designed and performed for a purpose, as shortly discussed in Box 2.2.

Box 2.2 – Why measure?

Any measurement is aimed at acquiring information on a property, and this sets its basic pragmatic context (as discussed for example in "Why measure?", a now classic paper by Charles West Churchman, 1959). While information is sometimes sought for purely knowledge-related purposes, in an encompassing perspective measurement is a key component of *information-enabled decision making* (Petri et al., 2021): in order to make appropriate decisions on an object, a rational procedure matches the desired state with the current state of the object, as it can be known by means of measurement. This suggests that the *general purposes* of measurements – and accordingly the general criteria for assessing the quality of measurement results – can be classified depending on whether (i) measurement is a process of information acquisition as such, or (ii) a process of information acquisition aimed at enabling some sort of decision making for some given purpose(s) in a given context, as happens when measurement results are used as inputs to a control system (for example in a decision rule of the kind: "if the value of the relevant property is less than X than do Y,

[14] The definitions of <measurement> – "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" (JCGM, 2012: 2.1) – and <measurement procedure> – "detailed description of a measurement according to one or more measurement principles and to a given measurement method, based on a measurement model and including any calculation to obtain a measurement result" (JCGM, 2012: 2.6) – given by the VIM are very clear in maintaining this distinction.

else do Z") or to set the parameters of predictive models.

An important case of this second class of purposes is *conformity assessment*, in which "a measurement result is used to decide if an item of interest conforms to a specified requirement" (JCGM, 2012b: p. vii), as when a supplier and a customer agree about the technical specifications of an object to be produced, and then the supplier has to decide whether a produced object is compliant with the specifications and therefore can be safely sent to the customer. Such specifications are typically stated in reference to one or more properties of the object, where for each property the target value – called the *nominal value* – is given (for example, the requirement that the mass of the object be 1.2345 kg) and then compared with the value that has been measured for that property. Only in the simplest, and in fact ideal, case, the decision rule is: "supply only if the measured value is the same as the nominal value". More realistically, some tolerance is usually allowed, so that "the requirement typically takes the form of one or two tolerance limits that define an interval of permissible values, called a tolerance interval, of a measurable property of the item", with the conclusion that "if the [measured] value of the property[, up to measurement uncertainty,] lies within the tolerance interval, it is said to be conforming, and non-conforming otherwise" (JCGM, 2012b: p. vii). Conformity assessment has widespread societal applications, as witnessed by the importance attributed to it in the "Blue Guide" on the implementation of EU product rules (2022/C 247/01[15]), where an entire section (5.1.1) is devoted to conformity assessment, defined as "the process carried out by the manufacturer of demonstrating whether specified requirements relating to a product have been fulfilled", with the condition that "a product is subjected to conformity assessment both during the design and production phase".

These two classes of purposes assume that measurement is a process aimed at describing the state of the object under measurement. In fact, when psychosocial properties are considered, a third class of purposes is possible, and actually exploited. If the object under measurement – in this case an individual or group of individuals – when informed of her/his/their condition of being under measurement, may change her/his/their state as a consequence, which can occur due to both conscious and unconscious processes, then the fact itself of being informed that a measurement is going to be performed may cause a change in the relevant property. A generalized version of these conditions is sometimes characterized as the *Hawthorne effect*, which "concerns research participation, the consequent awareness of being studied, and possible impact on behavior" (McCambridge et al., 2014: p. 267), an example of which is provided by "the audit culture of universities – their love affair with metrics, impact factors, citation statistics and rankings – [that] does not just incentivize [a] new form of bad behavior. It enables it." (Biagioli, 2016: p. 201). This third class of possible purposes of measurement leads to paradoxical situations in which communicating that a measurement will be performed, even without actually performing it, may be sufficient to obtain the expected outcome. In such cases measurement is not aimed at obtaining descriptive information on properties, but instead becomes a tool for inducing behavioral changes in individuals or groups of them.

For a number of possible reasons, the process may not fully implement what the procedure specifies. This justifies not only maintaining a clear distinction between the procedure and the process, but also adopting a systemic view and interpreting both of them as components of a system, that we

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022XC0629(04).

call a *measurement system*, designed, set up, and operated to solve a *measurement problem*, as depicted in Fig. 2.4.[16]

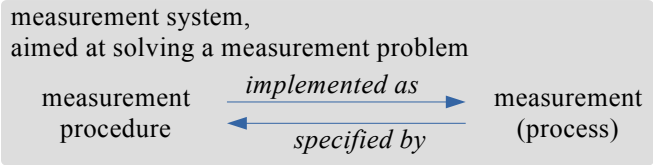


Fig. 2.4 Measurement systems, including the process of measuring and the procedure that specifies it

In summary, following the VIM, while measurement is an empirical process, the procedure that specifies it is an informational entity. This twofoldness – measurement systems including both empirical transformations and their specifications – is a characterizing feature of measurement, which is then a designed, not a natural and spontaneous, process.

Of course, this condition is still not sufficient: there are plenty of designed, input-output empirical processes that are not measurements. An example is a manufacturing process, in which raw materials are transformed into final products according to production plans. Clearly, a manufacturing process is not measurement. However, we propose that all examples of measurements can be uncontroversially characterized as designed processes.

In the case of the measurement of temperature, the empirical process is the interaction of the thermometer with the body of the individual under consideration; this interaction has to be performed according to a procedure that specifies in particular how the thermometer must be prepared and applied for guaranteeing that it is in appropriate thermal contact and for the appropriate duration with the body. In the case of the measurement of reading comprehension ability, the empirical process is the sequence of actions performed by the student in reading the text and answering the questions; this sequence has to be performed according to a procedure that specifies in particular how the student must be preliminarily instructed, how the environment must be prepared, the way the text passage and the test are delivered to the student, the time given to the student for completing the process, etc.

2.2.3 Measurement as a process whose input is a property of an object

The third black-box condition of measurement is that it requires an interaction with an *object* and that this interaction must be related to a *property* of that object, as depicted in Fig. 2.5.

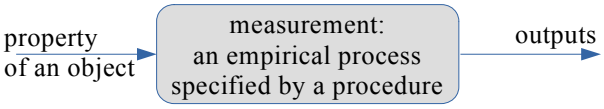


Fig. 2.5 The abstract structure of measurement (third version)

This assumes a basic ontology including *objects having properties*. For example, both rods and human beings are, in this sense, objects, and both have a temperature and a length (in the case of human beings, their height) among their properties; on the other hand, human beings, but not rods, have reading comprehension abilities.

[16] The concepts <measurement system> and <measurement problem> are not defined in the VIM. About the former, ASTM International, formerly known as American Society for Testing and Materials, developed a "Measurement Systems Analysis" (MSA) methodology (ASTM, 2022) according to which a measurement system is an encompassing entity, comprising all hardware and software devices, the procedures and methods involved in the process of measuring, together with the object that bears the property to be measured, its environment, and the human beings that operate in the process.

Fig. 2.6 proposes a pictorial representation of this simple ontology: the object *a* has a certain set of properties $P_1[a]$, $P_2[a]$, ...[17]

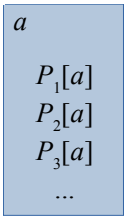


Fig. 2.6 A pictorial representation of the simple ontology we assume regarding objects and their properties: the object *a* has the properties $P_1[a]$, $P_2[a]$, $P_3[a]$, ...

This position – that measurement is necessarily of a property of an object – is less obvious than it might seem. For example, while Norman Campbell was clear on this matter in defining measurement as "the process of assigning numbers to represent qualities" (1920: p. 267),[18] the widely quoted rephrasing by Stanley Stevens – "measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules" (1946: p. 677) – may be interpreted as aimed at avoiding any reference to properties in a conceptual framework about measurement. While we devote part of Chap. 5 to this subject, we maintain here that, properly, we do not measure objects, but properties of objects. Warren Torgerson stated it clearly (1958: p. 14; the original text contains "system" where we have written "object"):

> While [the] distinction between [objects] and their properties is perhaps obvious, it is nevertheless an important distinction. It is of special importance here because of the fact that it is always the properties that are measured and not the [objects] themselves. Measurement is always measurement of a property and never measurement of a[n object].

Vice versa, some presentations claim that the input of a measurement is a value. For example, John Bentley (2005: p. 3) writes that "the input to the measurement system is the true value of the variable". If values of properties (and quantities) are distinguished from properties (and quantities) of objects, as we consider necessary to account for the basic conceptual structure of measurement (a point developed further in Chap. 5), the conclusion is that this is a categorical mistake: the empirical interaction on which a measurement is based cannot be with mathematical entities, such as property values.

Traditionally, measurement is presented as related to quantities, not more generically to properties. Given the importance of this subject, Chap. 6 discusses the relation between properties and quantities. We do not venture to propose a definition of what a property is, because it is a complex and controversial pre-metrological topic requiring a fundamental ontological framework (for an introduction see, e.g., Orilia & Swoyer, 2020). While Chaps. 5 and 6 are devoted to better analyzing this subject, it is sufficient to mention here that an empirical property of an object – and thus more specifically an empirical quantity of an object – such as the length of a rod or the reading comprehension ability of an individual, is associated with a *mode of empirical interaction* of the object with its environment. This association happens under the conditions that:

[17] The notation $P[a]$, with the symbol for the object delimited by square brackets, is used here to recall the functional notation, where in fact $P[a]$ stands for the property P of the object a and is not a mathematical function as such, but at the same time to emphasize that P can be formalized as a function.

[18] Note Campbell's use of the term "quality" in place of "property", which we avoid because <quality> explicitly contrasts with <quantity>: while stating that quantities are specific kinds of qualities is indeed odd, in the conceptual framework of the VIM – which we generally adopt here – quantities are specific kinds of properties. Furthermore, whether measurement is actually an assignment and its results are representations is an issue that we discuss in the following chapters.

- an object a empirically interacts with its environment in multiple modes, and each of them is supposed to correspond to a property of the object, say, its length $L[a]$, its weight $W[a]$, its reading comprehension ability $RCA[a]$, ...;
- some objects, a_1, a_2, ..., are comparable with respect to some of their properties, and sometimes distinct objects are discovered to have empirically indistinguishable properties, where indistinguishability, designated here as "≈" is weaker than equality (two properties could be indistinguishable by the available observational means, and nevertheless could be discovered to be different by adopting better tools); for example, a_1 and a_2 might be indistinguishable with respect to their length, $L[a_1] \approx L[a_2]$.

For example, a rod a_1 and a person a_2 can be compared with respect to their lengths, $L[a_1]$ and $L[a_2]$, and the length of the rod could match the height of the person, $L[a_1] \approx L[a_2]$, but it is not possible to compare the length of the rod, $L[a_1]$, with the reading comprehension of the person, $RCA[a_2]$, as schematically represented in Fig. 2.7.

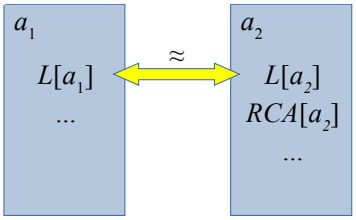


Fig. 2.7 The objects a_1 and a_2 can be compared with respect to their common property length, i.e., in principle $L[a_1] \approx L[a_2]$ is either true or false for a given comparison context, whereas $L[a_1]$ cannot be compared with the reading comprehension ability of a_2, $RCA[a_2]$, so that whether $L[a_1] \approx RCA[a_2]$ is meaningless

Such comparisons are empirical, not mathematical, as they involve empirical properties of objects. For example, assessing which of two objects is longer, or warmer, does not require operating with numbers, units of length, or temperature, etc. Hence, values of properties are still not needed at this stage. Similarly, although reading comprehension ability is most typically measured via a process that involves numbers, it is still the case that the reading comprehension ability of two individuals could be compared directly and without relying on values, for example by a judge asking questions to the two readers and then deciding which has the greater reading comprehension ability.

Comparable properties (or, by maintaining the explicit reference to the objects: properties relatively to which objects are comparable) are said to be *of the same kind* (JCGM, 2012: 1.2), so that the length of a rod and the height of a person are properties of the same kind, whereas the length of a rod and the reading comprehension ability of a person are not. The relational concept <kind of property> is reified, according to the principle that there exists an entity, length, of which both are instances, in the sense that both the length of the rod is a length and the height of the person is a length, and for which the two are comparable.

Usually the term "property" is used to designate both properties of objects and their kinds of properties, and the same happens for "quantity", so that it is said, for example, both that length is a quantity, and that the length of a rod is a quantity, being a quantitative property of the rod. Whenever avoiding this ambiguity is appropriate, we call an entity such as length or reading comprehension ability a *general property*, and an entity such as the length of a given rod or the reading

comprehension of a given person an *individual property*, and this would hold more specifically for quantities.[19]

The notation we adopt for general and individual properties is presented in Table 2.3.

Table 2.3 Notation for general and individual properties

The entity...	**... is designated as...**	**... and exemplified by...**
a general property	*P* (uppercase italic), so that P_i is the *i*th element of a set of general properties; hence for example the general property length is designated as *L*	*length, temperature, reading comprehension ability, ...*
an individual property	p (lowercase roman), so that p_i is the *i*th element of a set of individual properties that are instances of *P*; hence for example a set of lengths is designated as $\{\ell_1, \ell_2, ...\}$	*a given length, a given temperature, a given reading comprehension, ...*

[19] This is what a reference text (the so-called "Red Book") of the International Union of Pure and Applied Physics (IUPAP) says about the distinction between general and individual properties in the specific case of physical quantities: "There are two somewhat different meanings of the term physical quantity. One refers to the abstract metrological concept (e.g., length, mass, temperature), the other to a specific example of that concept (an attribute of a specific object or system: diameter of a steel cylinder, mass of the proton, critical temperature of water). Sometimes it is important to distinguish between the two and, ideally, it might be useful to be able to do so in all instances." (IUPAP, 2010: 1.1). Note that the terms that we adopt, "general property" and "individual property", are not standard, and, as already mentioned, usually the same term "property", and thus more specifically "quantity", is used to designate both general and individual properties. Other corresponding pairs of terms are "properties in the general sense" and "particular properties", as in the second edition of the VIM in reference to quantities (BIPM et al., 1993: 1.1), but also, e.g., "property" and "property manifestation" (Pfanzagl, 1971; Benoit & Foulloy, 2013), "attribute" and "level of attribute" (Michell, 2002), "quality" and "state of a quality" (Piotrowski. 1992), and – only applicable to quantities – "quantity" and "magnitude" (Holder, 1901, as translated by Michell and Ernst, and then adopted, among others, by Kyburg, 1997). Given that in some definitions of the VIM the term "magnitude" appears (e.g., 'quantity' is defined as "property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference", JCGM, 2012: 1.1), a few more words may be useful to justify why we do not use the term "magnitude" here. We have three basic reasons to justify this position. First, the term "magnitude" is used today with different and incompatible meanings – by claiming for example that magnitudes *are* quantities or that quantities *have* magnitudes (and in this second case the reference could be either to general quantities, e.g., mass has a magnitude, or to individual quantities, e.g., the mass of this object has a magnitude) – so that adopting it would require a more or less arbitrary selection. And while the term "magnitude" is used to translate the Greek μεγεθος, a lexical reference to this tradition is now outdated, given the Aristotelian contraposition of 'magnitude' and 'plurality' (πληθος, also translated as "multitude"): "a quantum is a plurality if it is numerable, a magnitude if it is a measurable" (Aristotle's Metaphysics, Book 5, Part 13). Second, the pair of terms "quantity" and "magnitude" seems to be so semantically superposed that in some languages other than English they are not distinguished (so that, for example the official French text of the VIM definition of 'quantity' mentioned above is: 'grandeur', "propriété d'un phénomène, d'un corps ou d'une substance, que l'on peut exprimer quantitativement sous forme d'un nombre et d'une référence": the concept 'magnitude' just disappeared...). The third reason is related to our interest in providing a general presentation of properties, of which quantities are a specific case. While the term "magnitude" could be intended as a synonym of "amount", so that for example one could say that mass is a quantity because objects have mass in amounts, non-quantitative properties do not have magnitudes (as in the VIM3 definition of nominal property, JCGM, 2012: 1.30), with the consequence that one or more terms corresponding to what magnitudes are for quantities should be adopted for non-quantitative properties. Indeed, sometimes for ordinal properties the term "level" is used to this goal. In summary, we believe that the pair "general property" and "individual property" provides a lexically simple and semantically encompassing terminology.

a property of a given object *a*	$P[a]$, so that $P[a_i]$ is the property P of the ith element of a set of objects; hence for example a set of lengths of objects a_1, a_2, ... is designated as $\{L[a_1], L[a_2], ...\}$	*the length of a given rod, the temperature of a given body, the reading comprehension of a given individual, ...*
a value of a property	p (lowercase italic), so that p_i is the ith element of a set of values of P; hence for example a set of values of length is designated as $\{\ell_1, \ell_2, ...\}$	1.23 m *as a value of length,* 2.34 °C *as a value of temperature, a reading comprehension ability of 1.23 logits on a particular RCA scale, ...*

A measuring system is designed to measure a whole set of instances of a general property and then to be applied to the measurement of properties of objects. The property of an object intended to be measured is called the *measurand* (JCGM, 2012: 2.3).[20]

A few words can be spent here also about objects: we are saying that what is measured is the property of a given object, but what is an object? Here we simply accept the pragmatic stance that an object is *anything that has properties*. However, this is not as trivial as it could seem. Consider the case of speed: as is well known, speed is a property that a body has only relatively to a frame of reference (i.e., speed is not an "absolute" property of a body). Hence, in this case the object under measurement is not the body alone, but the body together with the frame of reference.

The condition that measurement is a designed empirical process whose input is a property of an object is still not sufficient: there are plenty of such processes that are not measurements. An example is a transmission process, in which an input signal such as a stream of human voice is modulated into an electric potential difference that is then transferred to a channel; at the other end of the channel the process is reversed and the voice is obtained again. Clearly, a transmission process is not measurement, so we need a further condition.

2.2.4 Measurement as a property evaluation

The fourth black-box condition of measurement is that it produces *information* on the measurand *in the form of values of properties*, and thus, in the specific case of quantities, in the form of values of quantities. There is some confusion in the metrological literature about what values of quantities are, and the concept <value of a property> is seldom used, but entities such as 1.2345 m or 2.34 kg, are uncontroversially recognized as examples of values of quantities. Hence we introduce the subject here only for quantities with a unit, leaving to Chap. 6 the general treatment of also non-quantitative properties and their values.[21]

In the simplest case, in which measurement uncertainty can be omitted, a *measurement result* (JCGM, 2012: 2.9) is[22]

[20] For such a key concept the VIM unfortunately has only an entry about measurands as individual properties (e.g., the measurand is the length of rod *a*), but does not provide a term for the general property intended to be measured (e.g., length): we use "general measurand" in this case.

[21] Values of properties could be, for example, cube in a given set of shapes (a value of the nominal property *shape*), or second-preferred in a given sequence of preferences (a value of the ordinal property *preference*).

[22] As customary, we write this relation as an equality, =, instead of as an equivalence, ≅, or as a similarity, ≈. The nature of this relation is discussed in Chap. 5. More completely, a measurement result must include also information of some sort on the measurement uncertainty (JCGM, 2012: 2.26), a condition that in a following section we show to be a critical

measurand = measured value of a quantity

a relation that we call the *Basic Evaluation Equation* and whose meaning is analyzed in Chap. 5.[23] A symbolic form of a Basic Evaluation Equation is

$Q[a] = q_m$

For example

length of rod a = 1.2345 m

or in symbols

$L[a] = 1.2345$ m

as depicted in Fig. 2.8.

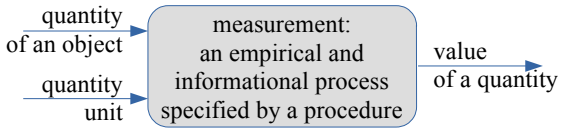


Fig. 2.8 The abstract structure of measurement (fourth version, in the case of quantities)

The relation can be then written more analytically as

$Q[a] = x\ q_{ref}$

where q_{ref} is the unit (JCGM, 2012: 1.9) and x is the numerical value of the quantity (JCGM, 2012: 1.20).[24] Such a relation is based on a direct or indirect comparison between two empirical entities (the quantity of an object and the chosen quantity unit) and reports it in terms of an information entity (the value of a quantity): that is why, as Norman Campbell famously put it, "the object of measurement is to enable the powerful weapon of mathematical analysis to be applied to the subject matter of science" (1920: p. 267). On the other hand, the actual nature of this relation is a matter of controversy: is it an assignment or a determination (Mari, 1997)? And more specifically: is it an attribution (of the value to the measurand), an expression (of the measurand by means of the value), a representation (of the measurand by means of the value), ... or indeed an equality? These questions are discussed in Chap. 5.

In any case, the information conveyed by a measurement result is acknowledged to be relational: it involves pairs of individual properties, which in the case of quantities are the measurand and a unit of the same kind as the measurand, and reports that the former is a given multiple of the latter, so that the previous equation may be rewritten in the form

$Q[a] / q_{ref} = x$

We call any process with this input-output characterization a *property evaluation*, or *evaluation* for short.[25]

Hence, measurement can be abstractly characterized as *a designed empirical property evaluation*. There is an essential difference between the input, which is empirical, and the output, which is

characteristic of measurement. Note that, together with "measured value", the GUM uses also the term "estimated value" (JCGM, 2008: 2.2.4), with a more explicit statistical-probabilistic connotation.

[23] The term "evaluation" inherits the ambiguity of "value", as mentioned in Footnote 7. We are using it here in the technical, non-axiological sense of *attribution of a value to the property of an object*.

[24] The notation q_{ref} for a generic unit is consistent with both the recommendations of the SI Brochure ("Unit symbols are printed in upright type regardless of the type used in the surrounding text. They are printed in lower-case letters unless they are derived from a proper name, in which case the first letter is a capital letter.", BIPM, 2019: 5.2) and our convention of designating individual properties with lowercase roman characters (about the nature of units as individual quantities, see Mari et al., 2018).

[25] Nordin et al. (2018) adopt the term "examination", which we consider less clearly referring to the production of values of properties.

informational: as already mentioned, a structural reason for the complexity of measurement is that it is neither purely empirical (like a physical transformation) nor purely informational (like a computation), but partly empirical and partly informational.

As noted above, the conditions presented so far are not specific to measurement, and other processes might fulfill them, like stating one's personal opinion in quantitative terms, by taking as input the property of an object and producing output consisting of one or more values that report the opinion of the subject who evaluates. Under what conditions measurement can be identified as a specific form of designed empirical property evaluation is a key issue for measurement science. The traditional position is to add constraints related to the structure of the measurable properties, and accept as measurements only the property evaluations whose inputs are quantitative properties (i.e., limiting the nature of the inputs and the outputs).[26] We discuss this standpoint first in Chap. 4 and then in Chap. 6. In Chap. 7 a different route is followed, by "opening the black box" in order to explore the concrete structure of the process and therefore to provide a justification of a relation that equates properties of objects and values, as in a Basic Evaluation Equation. From this analysis we derive some general conditions, which we propose to be sufficient to characterize measurement as a specific kind of designed empirical property evaluation.

Before concluding this chapter, one question remains to be discussed. According to our presentation, measurement is a process that starts from an *empirical* entity, i.e., a property of an object, and produces an *information* entity, in the simplest case the value of a property: how is this possible? Even a preliminary answer requires us to take a look at what is "inside the box".

2.3 Between the empirical world and the information world

The role of measurement as a process that connects entities of the empirical world and entities of the information world may be presented as in Fig. 2.9.[27]

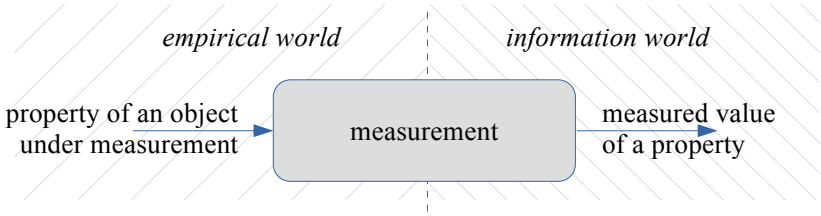


Fig. 2.9 Measurement as a process between the empirical world and the information world (first version)

Even in this introductory stage, a few words are appropriate to provide an explanation of this peculiarity. The mixed nature of measurement, partly empirical and partly informational, as mentioned in Sect. 2.2.1, originates from a process whose high-level, functional structure may be described as follows, in the structurally simplest case of what may be called a *direct measurement* (a short term for

[26] What defines a quantitative property is in turn a controversial issue. For example, while a strict interpretation assumes that properties must be additively composed to be quantities, sometimes (e.g., by Ellis, 1968: p. 25) only their linear ordering is required. Physical quantities, as length, duration, energy, etc., are examples of such quantitative properties.

[27] Here and in what follows we assume the distinction between entities of the empirical world and entities of the information world. While we do not dare to propose a general definition of what *empirical* and *informational* are and how they are related, a simple example may be helpful to convey the basic message, about a word and an utterance emitted by a speaker: the utterance is a physical phenomenon, and as such characterized by empirical properties such as its duration, frequency spectrum, and total energy; a word is instead a piece of information, characterized by its length (in number of characters), number (singular or plural), gender (masculine or feminine), and so on. Of course, attributing a gender to a sound or a bandwidth to a word is nonsense.

"measurement based on a direct method") (Giordani & Mari, 2019).[28]

1. *Transduction*. An empirical device – let us call it a *measuring instrument*[29]– is put in interaction with the object under measurement with respect to a property of the object, to which it is sensitive. The device operates as a transducer: as a result of the interaction, the device changes its state, by transducing the measured property to another property, usually called an *instrument indication* or "reading". For example: (a) the spring of a dynamometer, as is traditionally found in a bathroom scale, is intended to be a measuring instrument which transduces the applied weight force (the property being measured) to a spring elongation (the indication); or (b) a paper sheet with the printed text of a test comprising several multiple-choice items is intended to be a measuring instrument which transduces the reading comprehension ability of an individual to a response in the form of a pattern of marks on the printed checkboxes (the indication). Hence, this stage is entirely empirical.
2. *Instrument scale application*. A measuring instrument is designed so to make it possible to associate distinguishable indications with distinct information entities (typically but not necessarily numbers), which may be then called *indication values*,[30] through a mapping that is usually called a *scale* (for a more extended discussion of what scales are, see Box 6.2). For example: (a) the elongation of the spring is mapped to a value of length in a given unit; or (b) the pattern of marks on the printed checkboxes is mapped to a set of scored responses. Hence, this stage is partly empirical and partly informational.
3. *Calibration function computation*. The measurand and the indication are not, generally, properties of the same kind: hence, indication values are generally not accepted as suitable means to convey information about the measurand, as, in our examples, it would imply reporting information about a force by means of a value of length, and about a reading comprehension ability by means of a set of scored responses to test questions. The indication value is then mapped to a *measured value* by applying a function which models the transduction behavior of the instrument, and therefore the relation between the measurand and the indication, called the instrument *calibration function*.[31] For example, the value of length is mapped to a value of force, and the array of scored responses is mapped to a value of reading comprehension ability. Hence, this stage is entirely informational.

[28] In Sect. 3.2 we make the presentation more realistic by introducing measurement error / uncertainty; in Chap. 7 we present this as the basic structure of a direct method of measurement, and refine it in order to better identify its components.

[29] The VIM has different definitions for *measuring instrument*, a "device used for making measurements, alone or in conjunction with one or more supplementary devices" (JCGM, 2012: 3.1), and *measuring system*, a "set of one or more measuring instruments and often other devices, including any reagent and supply, assembled and adapted to give information used to generate measured quantity values within specified intervals for quantities of specified kinds" (JCGM, 2012: 3.2). The difference is subtle: is a balance a measuring instrument? its graduated scale? one of its pans? a screw in it? In order to maintain a distinction with a measurement system – as introduced in Sect. 2.2.2 – we will use one term "measuring instrument" for both. Hence a measurement system is an overall entity that includes both empirical and informational components, while a measuring instrument is an empirical component of a measurement system.

[30] The VIM defines <indication> as a value (JCGM, 2012: 4.1), but does not provide a term for the related property. Also for this reason (and consistently also with other JCGM documents, such as "The role of measurement uncertainty in conformity assessment", JCGM, 2012b), here "indication" refers to a property, whose values are called "indication values".

[31] More precisely, the function which models the transduction maps properties under measurement to indications. In performing this third stage it is then assumed that (i) through the instrument calibration the function is known and can be computed in terms of values of the involved properties, and that (ii) the function is invertible, so that values of the property under measurement can be obtained from values of the indication.

Hence, the sequence 1→2→3 starts in the empirical world from a measured property of an object and leads to a measured value in the information world, as depicted in Fig. 2.10.

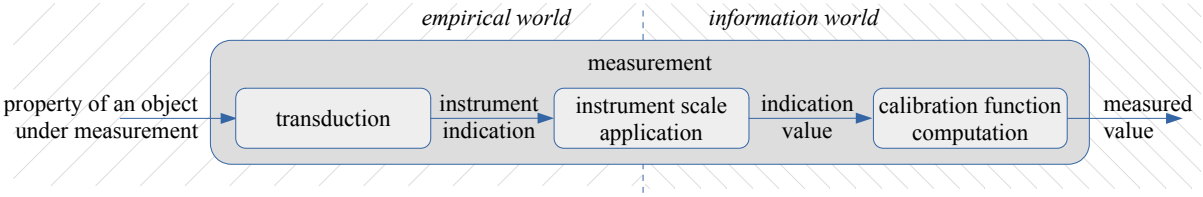


Fig. 2.10 Measurement as a process between the empirical world and the information world (second version)

In other words, the process produces an information entity, i.e., the measured value, which is expected to convey information about an empirical entity, i.e., the measured property, as summarized in a Basic Evaluation Equation

measured property of a given object = measured value

or more specifically measurand = measured value of a quantity, as discussed in Sect. 2.2.4. This shows the ambiguity of the term *measured property*, which at the same time may refer to the property

- which triggers the transduction, as performed by the measuring instrument, or
- to which the measured value is attributed.

While in simple or ideal cases the two coincide, in general the distinction is critical, and the property with which the measuring instrument interacts might not be the same as the one referred to in the Basic Evaluation Equation reporting the result of the measurement. The VIM defines <measurand> as "quantity intended to be measured" (JCGM, 2012: 2.3), thus making it clear that it refers to the property to which the measured value is attributed, and it uses the term "property being measured" (e.g., in JCGM, 2012: 2.3, Note 3) for the property which triggers the transduction.[32] With the aim of making this principled distinction as clear as possible, we use the term *intended property* for the measurand, and the term *effective property* for the property that produces an effect on the measuring instrument. In the case of direct measurement, the measurement procedure could be designed so as to make the measuring instrument interact exactly with the measurand, and therefore to make the effective property the same as the intended property. However in actual, empirical conditions the intended property and the effective property may be different, and such a difference might not be entirely under the control of the measurer: this is then a fundamental source of uncertainty that precedes the measurement as such. Given its structural dependence on the way the measurand is defined – see Box 2.3 – this has to do with what may be called *definitional uncertainty*, and it is the lower bound of measurement uncertainty, as discussed in Sect. 3.2.4.

Box 2.3 – Intended property and effective property

Let us discuss the distinction between intended properties and effective properties and their relations by means of an example (taken and developed from the GUM, JCGM, 2008: D.3.2-D.3.4). Suppose that we are interested in getting information on the thickness of a given sheet of material, being aware that, unavoidably, the sheet is not perfectly regular in its shape due to its slightly different thicknesses in different points, as known by a previous examination, possibly including some measurements, of the shape of the object. Depending on the reasons for which the information is to be acquired, at least four distinct strategies are possible for identifying / defining the measurand,

[32] This ambiguity affected the VIM itself, which in its first two editions defined <measurand> as "quantity subject to measurement" (ISO, 1984: 2.9; BIPM, 1993: 2.6), thus without a clear distinction between the two meanings.

each of which suggests the application of a measurement procedure, as related in particular to how the measuring instrument is applied, and is then about an effective property, i.e., the property with which the instrument should interact according to the measurement procedure (what the VIM calls "property being measured", e.g., in JCGM, 2012: 2.3, Note 3, and the GUM calls "quantity realized for measurement", e.g., in JCGM, 2008: D.2). Furthermore, each strategy of measurand definition corresponds to a definitional uncertainty, about the degree of specification provided by the definition, and can be characterized in terms of the transferability of measurement information that is produced. To ease the comparison, in what follows the four strategies are each given informal, evocative names and are presented in a schematic way, starting from the definition of the measurand and then in reference to these five aspects: (a) measurement procedure, (b) effective property, (c) definitional uncertainty, (d) measurement uncertainty, and (e) transferability of measurement information.

Strategy 1 (daily-life)

The measurand is defined as *the thickness of the sheet*, with no other specifications.

This is the strategy usually applied in daily situations in which some (possibly rough) information on the measurand is required for making decisions in the here-and-now conditions and nothing more.

(a) The instrument is applied once, possibly in a position considered representative of the thicknesses of the sheet. In principle, repeated application is possible, but not justified given the way the measurand is defined.

(b) The effective property is the thickness with which the instrument interacts when it is applied to the sheet in the environmental conditions in which it is applied.

(c) Definitional uncertainty is about the differences in thickness in different points of the sheet and the lack of specification of the environmental conditions.

(d) Measurement uncertainty depends on the accuracy of the instrument and the way it is applied.

(e) Measurement information is somewhat transferable, given the generic measurand model, as long as the possibly relatively large definitional uncertainty is considered acceptable.

Strategy 2 (operational)

The measurand is defined as *whatever thickness will be the input of the measuring instrument in the measurement in whatever environmental conditions there will be at the moment of the interaction between the instrument and the sheet*.

This is the purest operational strategy, and may be thought of as a refinement of Strategy 1, at the price of a complete lack of transferability of the information acquired in the process, which applies strictly only in the here-and-now conditions.

(a) The instrument is applied once, possibly even with no concerns about its position in the sheet and the environment condition in which it is applied. In principle, repeated application is not possible, given the way the measurand is defined.

(b) The effective property is, by definition, the same as the measurand.

(c) Definitional uncertainty is zero, given the identity of the measurand and the effective property.

(d) Measurement uncertainty depends on the accuracy of the instrument and the way it is applied.

(e) Measurement information is completely non-transferable.

Strategy 3 (*statistical*)
The measurand is defined as *the average thickness over the entire sheet in specified environmental conditions*.
This is a statistical strategy, that aims at making the produced information more transferable, given the specified environmental conditions.
(a) The instrument is applied in as many different points as the available resources allow, in order to reduce the risk of biases in the sampling. In principle, repeated application in each point is possible, to generate further statistical data.
(b) The effective property is the vector of thicknesses with which the instrument interacts when it is applied to the sheet in the environmental conditions in which it is applied.
(c) Definitional uncertainty is about the possibility that part of the sheet is not completely specified in the sampling, and therefore that the sample average is not considered of all relevant points of the sheet, and the incomplete specification of the environmental conditions.
(d) Measurement uncertainty depends on the accuracy of the instrument, the sampling conditions, and the correspondence between the specified and actual environmental conditions.
(e) Measurement information is transferable as long as the specified environmental conditions are met or a measurand model is available to make corrections.

Strategy 4 (*analytical*)
The measurand is defined as *the thickness in a specified point of the sheet in specified environmental conditions*.
This is the strategy that leads to the most specific information.
(a) The instrument can be applied multiple times, ideally always in the point specified by the measurand definition, to acquire information on the measurement uncertainty.
(b) The effective property is the thickness with which the instrument interacts when it is applied to the sheet in the environmental conditions in which it is applied.
(c) Definitional uncertainty is about the possibility that the point is not completely specified and the incomplete specification of the environmental conditions.
(d) Measurement uncertainty depends on the accuracy of the instrument, the correct application of the instrument to the specified point, and the correspondence between the specified and the actual environmental conditions.
(e) Measurement information is transferable as long as the specified environmental conditions are met or a measurand model is available to make corrections.

(Note that other strategies are possible, for example, as a variation of Strategy 3, if the measurand is defined as *the average thickness over the entire sheet in typical – in a sense to be specified – environmental conditions*.)

In the human sciences, one can see an example of the distinction between intended property and effective property in the case of measurement of reading comprehension ability. Here, the assessments always specify that the tests are to be given under conditions free from distraction while the student is reading the passages and responding to the comprehension questions, so that a noisy environment, for

example, would not be advisable. This is strongly associated with the intended property – a student's comprehension of text under good conditions. However, it may be the case that, in a given situation, a student is asked to respond in a noisy and distracting environment – this would be a case where the effective property differs from the intended property, and, presumably, any measurements made in this distracting situation would tend to show lower reading comprehension ability.

The understanding of measurement as a process that connects the empirical world and the information world is further generalized by acknowledging that values of properties – once obtained through measurement – are information entities that may be dealt with by mathematical means, such as the equations that formalize physical laws etc., so as to produce values of other properties, functionally related to those which have been measured by empirically interacting with them. A well-known example is about the density of an object, a value of which may be obtained by dividing the value of mass and the value of volume of the object, where the whole process is called *indirect measurement* (a short term for "measurement based on an indirect method"; in this case, density is measured indirectly via the direct measurement of mass and volume), as depicted in Fig. 2.11.

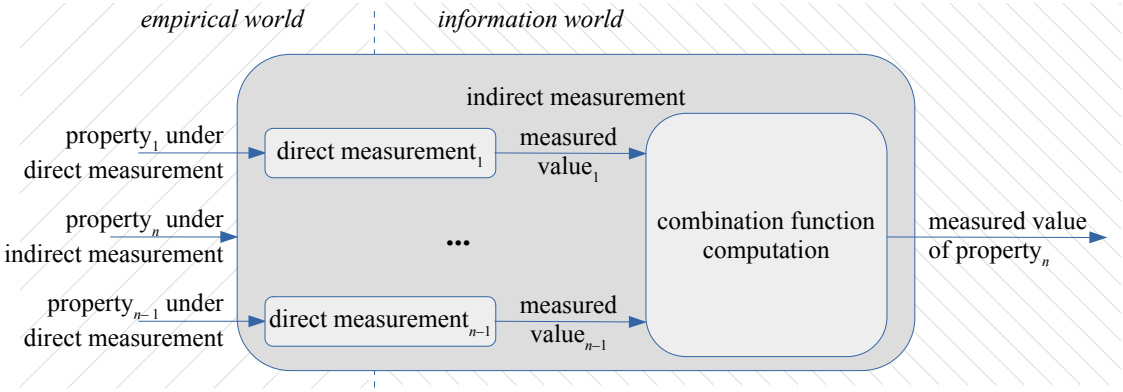


Fig. 2.11 Indirect measurement as a process including one or more direct measurements

In summary, according to this initial characterization *measurement is an empirical and informational process, designed on purpose, whose input is an empirical property of an object and that produces information in the form of values of that property*.

Though still only preliminary, this picture[33] is sufficient to furthering the development of our analysis. In the next chapter, we more systematically add to this the concept of uncertainty in measurement, and complete our account of what we see as fundamental concepts of measurement.

References

Anderson, P. W. (1972). More is different – Broken symmetry and the nature of the hierarchical structure of science. *Science, 177(4047)*, 393–396.

Aristotle (350 BCE). *Metaphysics*. Translated by W.D. Ross. Retrived from classics.mit.edu/Aristotle/metaphysics.html

ASTM International (2022). *ASTM E2782-17, Standard guide for Measurement Systems Analysis (MSA)*.

[33] The general idea that the measurement process is constituted of an empirical component and an information component is not new, of course. On this matter of particular interest are the presentations by Roman Morawski, who introduces the two components as *conversion* and *reconstruction* (2013), and by Giovanni Battista Rossi and Francesco Crenna, who call them *observation* and *restitution* (2018).

Benoit, E., & Foulloy, L. (2013), The role of fuzzy scales in measurement theory. *Measurement, 46*, 2921–2926.

Bentley, J. P. (2005). *Principles of measurement systems* (4th ed.). New York: Pearson.

Biagioli, M. (2016). Watch out for cheats in citation game. *Nature, 535*, 201.

Bridgman, P. W. (1959). How much rigor is possible in physics? In L. Henkin, P. Suppes, & A. Tarski (Eds.). *The axiomatic method* (pp.225-237). Amsterdam: North-Holland.

Bunge, M. (1974). *Treatise on basic philosophy – Vol.1 – Semantics I: Sense and reference*. Dordrecht: Reidel.

Bunge, M. (1983). *Treatise on basic philosophy – Vol. 6, Epistemology & Methodology II: Understanding the world*. Dordrecht: Reidel.

Campbell, N. R. (1920). *Physics – The elements*. Cambridge: Cambridge University Press.

Churchman, C. W. (1959). Why measure?. In C. West Churchman, P. Ratoosh (Eds.), *Measurement: definitions and theories*. New York: Wiley.

Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

Ellis, B. (1968). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

European Union (2014). *Directive 2014/32/EU of 26 February 2014 "on the harmonisation of the laws of the Member States relating to the making available on the market of measuring instruments"*. Retrieved from ec.europa.eu/growth/single-market/european-standards/harmonised-standards/measuring-instruments_en

Giordani, A., & Mari, L. (2019). A structural model of direct measurement. *Measurement, 145*, 535–550.

Hempel, C. G. (1966). *Philosophy of natural science*. New York: Prentice-Hall.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte uber die Verhandlungen der Koeniglich Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physikaliche Klasse, 53*, 1–46. Part 1 translated in Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement. Journal of Mathematical Psychology, 40(3), 235–252.

International Bureau of Weights and Measures (2019). *The International System of Units (SI) ("SI Brochure")* (9th ed.). Sèvres: BIPM.

International Bureau of Weights and Measures (BIPM) and other six International Organizations (1993). *International Vocabulary of Basic and General Terms in Metrology (VIM)* (2nd ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML), International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP).

International Organization for Standardization (2019). *ISO 1087:2019, Terminology work – Vocabulary*. Geneva: ISO.

International Organization for Standardization (2022). *ISO 704:2022, Terminology work – Principles and methods* (4th ed). Geneva: ISO.

International Organization of Legal Metrology (2022). *OIML V1:2022, International vocabulary of terms in legal metrology (VIML)*. Paris: OIML. Retrieved from www.oiml.org/en/files/pdf_v/v001-ef22.pdf; online version: viml.oiml.info

International Union of Pure and Applied Physics (2010). *IUPAP: SUNAMCO 87-1, Symbols, units, nomenclature and fundamental constants in physics ("Red Book")*. IUPAP, 1987 revision (2010 reprint).

Joint Committee for Guides in Metrology (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2009). *JCGM Charter*. Sèvres: JCGM. Retrieved from www.bipm.org/documents/20126/2071204/JCGM+Charter.pdf/298fcd3e-a464-f7e4-045f-8bd35429ee90

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.). Sèvres: JCGM (2008 version with minor corrections). Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2012). *JCGM 106:2012, Evaluation of measurement data – The role of measurement uncertainty in conformity assessment*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Kyburg Jr, H.E. (1997). Quantities, magnitudes, and numbers, *Philosophy of Science, 64(3)*, 377–410.

Lockhart, P. (2012). *Measurement*. Cambridge: Belknap.

Mari, L. (1997). The role of determination and assignment in measurement. *Measurement, 21*, 79–90.

Mari, L. (2013). A quest for the definition of measurement. *Measurement, 46*, 2889–2895.

Mari, L., Ehrlich, C. D., & Pendrill, L. R. (2018). Measurement units as quantities of objects or values of quantities: a discussion. *Metrologia, 55*, 716–721.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Mari, L., & Ruffini, R. (2018). An analysis of Goodhart's law toward a shared conceptual framework of measurement across the sciences. *Journal of Physics - Conference Series, 1065 072022*.

McCambridge, J., Witton, J., & Elbournec, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67(3)*, 267–277. Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC3969247

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54(2)*, 99–104.

Morawski, R. Z. (2013). An application-oriented mathematical meta-model of measurement. *Measurement, 46*, 3753–3765.

Nordin, G., Dybkaer, R., Forsum, U., Fuentes-Arderiu, X., & Pontet. F., Vocabulary on nominal property, examination, and related concepts for clinical laboratory sciences (IFCC-IUPAC Recommendations 2017). *Pure and Applied Chemistry, 90(5)*, 913–935.

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. Harcourt: Brace & World.

Orilia, F., & Swoyer, C. (2020). Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/propertiesPearl, J. (2009). *Causality – Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Pfanzagl, J. (1971). *Theory of measurement*. Berlin: Springer.

Piotrowski J. (1992). *Theory of physical and technical measurement*. Amsterdam: Elsevier.

Rossi, G. B., & Crenna, F. (2018). A formal theory of the measurement system. *Measurement, 116*, 644–651.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103(2684*), 677–680.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Chapter 3.
Technical and cultural contexts for measurement systems

This chapter aims to outline the technical and cultural contexts in which measurement systems, as presented in the previous chapter, are designed, set up, and operated. It first introduces the basic proposal that a measurement should produce as result not only one or more values of the property under consideration but also some information on the quality of those values, and discusses the consequences in terms of measurement uncertainty. This proposal is then embedded in the broader context of metrological systems, which help justify the societal significance of measurement results via their traceability to conventionally-defined measurement units, so that measurement results can be interpreted in the same way by different persons in different places and times. Finally, we consider the issue of what is measured, i.e., the property of a given object, or *measurand*, which must be somehow defined and identified. On this basis, the chapters that follow develop and bring further specificity to our analysis and proposals. As with the previous chapter, we believe that the contents of this chapter should be sufficiently uncontroversial to be read and accepted by most, if not all, researchers and practitioners.

3.1 Introduction

As presented in the previous chapter, the characterization of measurement as a process specified by a procedure is given in terms of a set of necessary conditions. This description is so fundamental that it provides a very abstract picture of what measurement actually is, and so, it now needs to be expanded and placed in a more concrete context. This is the purpose of the present chapter, which develops along three parallel lines.

The first line of development starts from the acknowledgment that the empirical nature of the properties to be measured and of the process of measuring them prevents exact (i.e., perfectly certain and specific) evaluations. On this basis, we elaborate the requirement that measurement should produce information about both values of the measurand and the quality of such an evaluation. This examination of measurement quality must begin with the features of the measuring instrument itself: the empirical nature of the instrument implies unavoidably non-ideal behavior that affects the quality of the results it generates: the lack of complete accuracy of the instrument is a key (though not the only) source of the errors and the uncertainties in the results.

The second line of development starts from the acknowledgment that measurement is a relational process – in the usual case for physical quantities it compares the measurand and the chosen unit – and embeds it into the scientific, technical, and organizational system that is the foundation for this relationship. This is made possible through the definition and the dissemination of measurement standards embodying the relevant reference properties. Measuring instruments are calibrated by means of such standards.

The third line of development complements this operational context with a conceptual one, by tracing from a historical perspective the common position that only quantitative properties are

measurable. The outcome of our analysis is that the Euclidean basis of this assumption is not sufficient as such for maintaining the constraint: this will pave the way for a further analysis of measurability.

3.2 The quality of measurement and its results

In Chap. 2 we introduced the distinction between a measurement procedure and a measurement process: the former is the description that specifies how the latter, i.e., the process, must be performed. Even if the specifications are exactly fulfilled, two measurement processes implementing the same procedure on the same object may produce different results. This calls for an explanation.

In principle, two situations might obtain. The property with which the measuring instrument is designed to interact either

- *has changed*, so that different measurement results may correctly report the fact that the object under measurement modified its state in the interval between the interactions, or
- *has not changed*, but the behavior of the measuring instrument has been affected by changes in the state of the environment or of the measuring instrument itself, so that different measurement results incorrectly report a difference that is not related to the measurand.[1]

From the perspective of measurement as such, the first case is not problematic: the property under measurement may actually change as the result of its dependence on other properties, of the object or the environment. We call any property whose changes produce a change in the measured property an *affecting property*. For example, if the measured property is the length of an iron rod, then, due to thermal expansion, the temperature of the environment is an example of an affecting property. If the measured property is the reading comprehension ability of a student, an example of an affecting property might be the intensity of distracting noises from the environment, insofar as such noise could negatively affect the student's ability to comprehend a text they are attempting to read. However, there can be properties other than the measurand which alter the behavior of the measuring instrument and therefore generate the second case; these are called *influence properties* (JCGM, 2012: 2:52). In the example of the measurement of the thickness of a rod by means of a caliper, an example of an influence property is the parallelism of the jaws, whereas, in the example of the measurement of the reading comprehension ability of a student by means of a test with multiple choice questions, an example of an influence property might be human error in the scoring of the item responses. The role of affecting properties and influence properties in a measurement is depicted in Fig. 3.1.

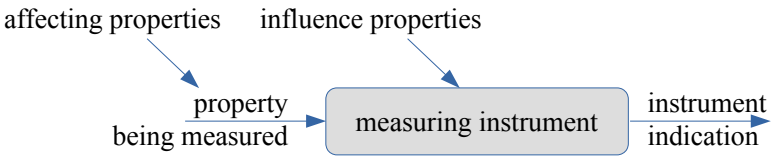


Fig. 3.1 A black box model of the empirical behavior of a measuring instrument

In a typical context, the two discussed situations are not mutually exclusive, and in fact they might co-occur to generate the multiplicity of results mentioned above. Under the principled hypothesis that the

[1] Eran Tal (2019) builds upon this distinction and argues that "due to the possibility of systematic error, the choice between [these two situations] is underdetermined in principle by any possible evidence": we do not further develop his argument here. Moreover, we note here that there is a third case: whether the individual property did change or not, what may change over time is *the definition of the general property* of which the individual property is an instance, thus possibly making the measuring instrument inadequate. While this is now unusual for physical properties, this situation is still not uncommon in the human sciences, as for example for nursing ability, the very definition of which depends on the cultural context and therefore changes over places and times. We discuss the problem of the existence and identification of general properties in Sect. 6.6.

influences described in the second situation can be identified, the problem arises of how to deal with the incorrect results that are obtained. Two basic strategies can be envisaged.

An *empirical* strategy aims at improving the behavior of the measuring instrument by reducing its sensitivity to the influence properties, and therefore the misleading variability of its results. This is a positive outcome, generally obtained at the price of additional resources (including money, competencies, time, etc.) devoted to the measurement. Of course, this is not always feasible.

The fact that measurement is both an empirical and an informational process makes possible a complementary, *informational* strategy: if the undesired variability cannot be completely removed, it can at least be modeled, evaluated, and formally expressed. The fundamental outcome is the acknowledgment that only in the simplest cases is the information acquired on a measurand by means of a measurement entirely conveyed by a single measured value. Generally, a structurally more complex result has to be reported instead. This is why the *International Vocabulary of Metrology* (VIM) defines <measurement result> as a "set of quantity values being attributed to a measurand together with any other available relevant information" (JCGM, 2012: 2.9). This complexity is justified under the assumption that "when reporting the result of a measurement [...], it is obligatory that some quantitative indication of the *quality* of the result be given so that those who use it can assess its reliability"[2] (JCGM, 2008: 0.1; emphasis added). In fact, one may even take as a definitional condition of measurement that its results include some information about their quality, as proposed for example by the U.S. National Institute of Standards and Technology (NIST) (Possolo, 2015: p. 12).

The subject is intermingled with the development of statistics and the theory of probability (see, e.g., Hacking, 1975; 1990, and Rossi, 2014), and the diversity of interpretations of probability is reflected in the diversity of understandings of the role of probability in measurement. In particular, according to the VIM, there are two "philosophies and descriptions of measurement", identified as the "Error Approach (sometimes called Traditional Approach or True Value Approach)" and the "Uncertainty Approach" (JCGM, 2012: Introduction). While for some purposes this might be too rough a classification, and some cases may be intermediate (see, e.g., Giordani & Mari, 2014), or even of a different kind (see, e.g., Ferrero & Salicone, 2006), we adopt this distinction to introduce the informational strategy and therefore the pivotal concepts of *measurement error* and *measurement uncertainty*.[3] However, before discussing such concepts, we first introduce a framework in which they may be understood.

3.2.1 A sketch of the framework

"Measurement is essentially a production process, the product being numbers." (Speitel, 1992). While we adopt a more encompassing view with respect to what is produced by a measurement (that is, not only numbers), we agree that measurement is a production process, and as such the quality of the products and the quality of the process are in principle distinct (though plausibly related), and as such they each deserve some consideration. It is clear that users focus on the quality of what is produced: in general, users would like to have trustworthy, useful information on the measurands in which they are interested. Were it possible to disentangle the quality of the process from the quality of the products, the former would become immaterial. But, as with any production process, the quality of

[2] In practice, this is possibly one of the sharpest distinctions between measurement in scientific and non-scientific contexts. Even the VIM admits that sometimes "the measurement result may be expressed as a single measured quantity value" and then acknowledges that "in many fields, this is the common way of expressing a measurement result" (JCGM, 2012: 2.9, Note 2).

[3] Like most of the contents of this chapter, what follows generally applies to both quantitative and non-quantitative properties, even though the mathematical aspects are mainly introduced here in reference to quantities. The issue of uncertainty in non-quantitative evaluations is further considered in Chap. 6.

the products – measurement results in this case – depends on the quality of the process, which is why both need to be taken into account.

The quality of a process of measurement has to do with the features of the experimental setup, which includes the measuring instrument(s) and everything that is exploited to control the environment with the aim of reducing its effects on the behavior of the instrument(s), so that in the best case the output of the instrument conveys information only about the measurand, and nothing else. As mentioned in Sect. 2.3, measurement is sometimes modeled as a black box that transduces an input property, i.e., the property being measured, to an output property, i.e., the instrument indication, under the acknowledgment that the transformation is usually affected by some influence properties. Such a transformation is modeled as an empirical transduction function, whose informational inverse is the instrument calibration function, which maps values of the instrument indication to values of the measurand.

On this basis a wealth of models and accompanying parameters have been developed *to characterize the behavior of measuring instruments*, where the preliminary distinction needs to be maintained between the "nominal" behavior of an instrument, as shown in the controlled conditions of instrument manufacturer laboratories, and its "field" behavior in actual measurements, the former typically being the limit, best case of the latter. The characterization of the nominal behavior of an instrument is supposed to be performed in such a way that the values of the relevant, both input and output, properties are known independently of the instrument under test, and this makes it clear that such a characterization is not a measurement, though it may exploit previous measurement results. In what follows the values of the property being measured are designated as x_i, the corresponding values of the instrument indication as y_j (where both the property being measured and the instrument indication are assumed to be scalar for the sake of simplicity), and the values of the (vector of the) influence properties affecting the indication as $\boldsymbol{z}_k$, so that the transduction performed by the instrument is modeled as a function f of the form $y_j = f(x_i, \boldsymbol{z}_k)$, that is, qualitatively, *output* = f(*sought input*, *spurious input*). Among the several parameters that characterize the behavior of a measuring instrument, and therefore its transduction function f, we focus on:

- *sensitivity*, the ability of instruments to produce distinct outputs in response to distinct expected inputs;
- *selectivity*, the ability of instruments to produce outputs unaffected by spurious inputs;
- *stability*, the ability of instruments to produce the same outputs in response to the same expected inputs even in different time instants;
- *resolution*, the ability of instruments to detect distinct expected inputs.

Together with other parameters not mentioned below, like discrimination threshold (JCGM, 2012: 4.16) and dead band (JCGM, 2012: 4.17), these "low level" features of measuring instrument contribute to the "high level" features of *precision*, *trueness*, and finally *accuracy*, as discussed in the rest of this section. It is remarkable that in principle these parameters are structural features of measuring instruments, and as such in principle independent of whether what the instruments measure are physical or psychosocial properties, and whether such properties are quantitative or not. As a matter of fact, the tradition of physical instrumentation for quantitative properties is far more developed on this subject, and that is why we provide here definitions that apply to quantities, by thus including differences, ratios, etc., with a simple running example of a spring dynamometer, which transduces the input weight force (the property being measured) to an output spring elongation (the instrument indication) (extensions to non-quantitative properties are possible indeed, as discussed in Mencattini and Mari, 2015).

Despite the formal possibility of these parameters being used in psychosocial measurements, in fact they are not common in the related literature. The main reason is the requirement that the values of the input property need to be known, whereas there are no "controlled conditions of instrument manufacturer laboratories", as was noted above, for most psychosocial contexts. Nevertheless, these parameters are indeed sensible ones and would be useful if available, even in somewhat approximate and/or different guises, and hence, we comment on such possibilities in the paragraphs below. We use somewhat different specific examples for different parameters, as distinct possibilities arise more commonly in some contexts than others, but will nevertheless try to incorporate the running example of reading comprehension ability (RCA) which has been introduced earlier.

The *sensitivity* of a measuring instrument, according to the VIM (JCGM, 2012: 4.12, adapted), is the "quotient of the change in an indication value of a measuring instrument and the corresponding change in a value of a property being measured", under the supposition that the influence properties remained constant; hence

$$sens(instrument, x, z) = [f(x+\Delta x, z) - f(x, z)] / \Delta x = \Delta y_z / \Delta x$$

for a sufficiently small change Δx. The sensitivity of the spring may be a function of the value x of the property being measured in the appropriate unit (e.g., metres per newton in the SI), describing how much the spring elongates in response to a change of the applied force, while all influence properties remain constant, and may be in turn also a function of such constant values of the influence properties. An instrument such that the indication value y depends linearly on the measured value x has constant sensitivity, and an instrument whose sensitivity is zero for a given set of forces is useless for measuring a force in that set. However, the sensitivity of an instrument may be different for different measured values, and this is modeled with a sensitivity function, that for some physical transducers is logarithmic, with decreasing sensitivity as the measured value increases.

In the context of psychosocial measurements,[4] sensitivity can be qualitatively ascertained up to a limit. In a situation where there is a broadly accepted and recognized difference between, say, two readers in terms of their RCAs, then one could ascertain the ability of a certain RCA test to distinguish them, say, by an expert panel of judges of readers' RCAs. This qualitative observation could then be used in a more controlled way to establish an ordinal categorization of multiple such judgments, such as for say, typical grade 2 students, typical grade 3 students, etc. This could then be used to describe, again in a qualitative way, the sensitivity of a certain RCA test to distinguish students with those RCAs. Thus, the logic pertains in the psychosocial setting, even if the realization of the parameter is not precise.

The *selectivity* of a measuring instrument is basically its insensitivity to influence properties; while the property being measured is constant, the selectivity of an instrument with respect to a given influence property can be evaluated as

$$sel(instrument, z, x) = \Delta z / [f(x, z+\Delta z) - f(x, z)]$$

for a sufficiently small change Δz. The selectivity of the spring with respect to temperature is a value in the appropriate unit (kelvin per metre in the SI), describing how much the spring elongates in response to a change in the environmental temperature, while the applied force and all other influence

[4] There is use of the term "sensitivity" in some areas that are mainly centered on health studies. In those contexts, "sensitivity", which is also called the *true positive rate* in these contexts, is defined as the proportion of actual positives which are correctly identified as such – for example, the percentage of sick people who are correctly identified by the test as having the relevant condition (e.g., Altman & Bland, 1994). This is a different concept, and not to be confused with the use of the term here.

properties remain constant: the lower the selectivity, the better the instrument.[5]

An example of an influence property for a mathematics test is the RCA of the student taking the mathematics test. Thus, the selectivity of the test with respect to the RCA of the reader is the ratio of the change in the test outcomes for a given change in the RCA of the reader. As noted above, this ratio may vary depending on the value of the reader's mathematics knowledge, and hence an estimate of the derivative with respect to RCA may be a selectivity function.

The *stability* of a measuring instrument is basically its insensitivity to time, thus under the acknowledgment that f may depend on time; while both the property being measured and the influence properties are constant, the stability of an instrument can be evaluated as

$$stab(instrument, \Delta t, x, z) = \Delta t / [f(x, z, t+\Delta t) - f(x, z, t)]$$

for a sufficiently small change Δt. The stability of the spring is a value in the appropriate unit (seconds per metre in the SI), describing how much the spring changes its elongation in different time instants, while the applied force and all influence properties do not change (stability considered in short intervals of time is also called *repeatability*): the greater the stability the better the instrument.[6]

In theory, for an RCA test, the stability could be observed by re-testing the students after a suitably small period. In the case of RCA, as with many psychosocial tests, it is unlikely that stability could be evaluated with a single test, as the readers will almost certainly recognize the passages and the RCA questions, and then recall their own answers – and hence the property under measurement would be confounded with readers' memories of the first test, and so this "theoretical" approach is not useful. However, an example of an instrument that would be relevant here would be a t*est bank*, where the stability of RCAs over different choices from the bank could be observed, and students can be observed using different tests, which have no overlapping items, from the test bank.

The *resolution* of a measuring instrument, according to the VIM (JCGM, 2012: 4.14, adapted), is the "smallest change in a property being measured that causes a perceptible change in the corresponding indication", thus again under the supposition that the influence properties remained constant; hence

$$res(instrument, x, z) = \min(\Delta x), \text{ such that } f(x+\Delta x, z) - f(x, z) = \Delta y \neq 0$$

The resolution of the spring is a value in the appropriate unit (newton in the SI), describing the minimum change of applied force that changes the spring elongation, while all influence properties remain constant: the smaller the resolution the better is the instrument.

In the case of psychosocial measurements, the difficulty of obtaining the values x of the property, as noted above, makes this parameter difficult to calculate. Nevertheless, some insight into the resolution can be found, based on the discrete nature of the items in the instrument. For a RCA test, the minimum detectable change in the indication would be a change where one of the set of items changed from being incorrect to correct (assuming that the items were dichotomous). This amount (i.e., the change) would depend on (a) the RCA[7] of the reader, and (b) (for some situations) which item changed. As for the general definition, the resolution would then be the minimum of these values.

[5] The definition in the VIM (JCGM, 2012: 4.13) takes into account all influence properties at the same time, and therefore remains qualitative. Moreover, the VIM itself notes (4.13, Note 1) that in some contexts selectivity is considered as insensitivity to influence properties of the same kind as the measurand.

[6] The VIM definition is very general: "property of a measuring instrument, whereby its metrological properties remain constant in time" (JCGM, 2012: 4.19).

[7] In practice, this would have to be measured.

What follows is a summary characterization of these lower-level parameters in terms of two higher-level parameters, which are illustrated in Fig. 3.2.

The *precision* of a measuring instrument (by adapting the VIM: JCGM, 2012: 2.15, modified in reference to ISO, 1994: 3.12) is the closeness of agreement between indication values or measured values obtained by repeated independent measurements on the same or similar objects under specified conditions. If measurement is modeled as affected by errors (see Sect. 3.2.2), precision is inversely related to the random component of errors, i.e., the one that is reduced by increasing the size of the sample of values. Hence, precision is evaluated by means of statistics of dispersion, like standard deviation, and, if evaluated on samples of indication values, does not require the instrument to be calibrated. In fact, the precision of a measuring instrument may be effectively estimated by the precision of its results, obtained in test conditions. In psychosocial measurement, precision is usually termed "reliability", which is evaluated as a degree, from 0 to 1: for an RCA test, precision would typically be estimated by one or more reliability coefficients, including internal consistency reliability (consistency of the results across the items within the test), test-retest reliability (consistency of a test results across different administrations), and inter-rater reliability (for item-formats that require human judgments, consistency of different raters, either at the item or the test level).

The *trueness* of a measuring instrument (by adapting the VIM: JCGM, 2012: 2.14, modified in reference to ISO, 1994: 3.7) is the closeness of agreement between the average value of a large series of measured values, obtained by replicate independent measurements on the same or similar objects under specified conditions, and an accepted reference value. If measurement is modeled as affected by errors, trueness is inversely related to the systematic (i.e., non-random) component of errors, i.e., the one that does not depend on the size of the sample of values. The trueness of a measuring instrument may be effectively evaluated as the inverse of measurement bias (JCGM, 2012: 2.18), with respect to a reference value taken from an available calibrated measurement standard in the process of metrological confirmation of the instrument (JCGM, 2012: 2.44).[8] Trueness is not commonly used in the human sciences, though it bears some similarity to early definitions of validity, discussed further in Sect. 4.3.

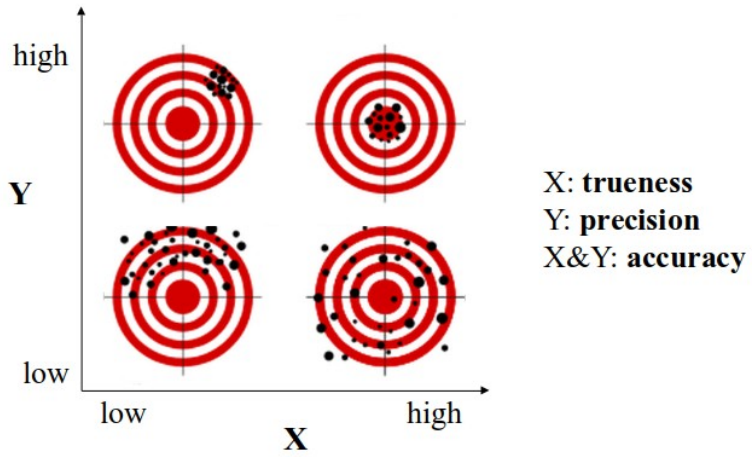


Fig. 3.2 A visual metaphor for precision, trueness, and accuracy; note that precision is independent of the presence of the bullseye, thus emphasizing that precision can be evaluated even if no reference values are known and the instrument is not calibrated, while trueness and then accuracy cannot

[8] When a measuring instrument is used in measurement, there are no known reference values to be applicable, as trueness should be evaluated with respect to the (unknown, if even existing) true value of the measurand. Hence the trueness of measurement results "is not a quantity" (JCGM, 2012: 2.14, Note 1).

As defined, precision and trueness are complementary features of an instrument: when the property being measured does not change, precision is about the capability of the instrument to maintain the values it produces to be close to one another, and trueness is about the capability of the instrument of maintaining the values it produces to be on average close to the target reference value.

Of course, a measuring instrument is expected to have at the same time a good precision and a good trueness. By combining precision and trueness, the highest-level parameter for characterizing the quality of a measuring instrument is obtained:

The *accuracy* of a measuring instrument (by adapting the VIM: JCGM, 2012: 2.13, modified in reference to ISO, 1994: 3.6) is the closeness of agreement between a measured value and an accepted reference value. If measurement is modeled as affected by errors, accuracy is inversely related to errors. The accuracy of a measuring instrument is evaluated in test conditions, by somehow combining its precision and trueness with respect to a reference value taken from an available calibrated measurement standard in the process of *metrological confirmation* of the instrument.[9] An accurate instrument is exactly what we would like to have: an instrument that produces trustworthy values.

As is discussed at further length in Sect. 4.3, in the human sciences one encounters the term "validity". In its earliest usages, which are still common in some areas of the literature, validity refers to closeness of agreement between measured values and a true value, sometimes operationalized in terms of an accepted reference value, as discussed in Sect. 4.3.1. Thus, from this perspective, the early version of validity is similar to accuracy.

Given this sketch of a framework about the characterization of the quality of the process of measuring, the issue arises of whether the quality of measurement results is entirely and solely determined by the quality of the measurement that produced them.

3.2.2 The Error Approach (or: True Value Approach)

Appropriate applications of the empirical strategy mentioned in the opening of Sect. 3.2 generally result in improvements of the behavior of measuring instruments, and consequently reductions in the variability of their results. Extrapolating from this, one might suppose that, if this improvement process were to continue indefinitely, a single, definite value would actually be obtained: "by analyzing the meaning of the obtained results of measurement, the experimenter ponders on the *true value*, the value that the best possible instrument would have generated" (translated from Idrac, 1960, emphasis added). Indeed, this hypothesis has a simple statistical basis. To see this, assume that the measurement is repeatable – that is, that (a) the environment and the measuring instrument are sufficiently stable, and that (b) the observed variability between interactions is only due to the influence of small, independent causes. Then, the repeated interaction of the instrument with the measurand generates a sample of values whose distribution becomes more and more stable as the number of values increases. More specifically, if the values can be averaged in a meaningful way (hence this obviously does not apply to nominal and ordinal properties), and s is the sample standard deviation, it is well known that the standard deviation of the distribution of the sample mean (the so-called "standard error") is estimated by $s/\sqrt{n}$, where n is the sample size. By increasing the sample

[9] What is commented above about the trueness of measurement results also applies to accuracy: when a measuring instrument is used in measurement, the accuracy of measurement results "is not a quantity" (JCGM, 2012: 2.13, Note 1). Vice versa, the instrument manufacturer operating in laboratory conditions may characterize, also quantitatively, the accuracy of the instrument, sometimes reported as its membership to an *accuracy class*, the "class of measuring instruments or measuring systems that meet stated metrological requirements that are intended to keep measurement errors or instrumental measurement uncertainties within specified limits under specified operating conditions" (JCGM, 2012: 4.25).

size, i.e., repeating the number of interactions of the measuring instrument with the measurand, the variability of the sample mean converges to zero, showing that the total influence of the aforementioned "small causes" is progressively reduced. Reasonably, then, the mean of a sufficiently big sample estimates "the mean value that the best possible instrument would have generated".

However, in order to make this concept of "best possible instrument" operational, a second condition must be fulfilled: an instrument is expected to maintain its calibration over time, so that, together with the calibration information, its indication values are sufficient to obtain appropriate values of the measurand. If the calibration information were not updated, the measurement results produced by the no-longer-calibrated instrument would be systematically biased and, critically, this bias would not be revealed by the repeated application of the instrument itself, as it is unaffected by sample size. In other words, the convergence to a target distribution is necessary but not sufficient to obtain the value that would be generated by "the best possible instrument".

This is a delicate point: the quality of the information produced by measurement is hindered by causes traditionally treated as belonging to two distinct and independent kinds, called *random* and *systematic* respectively. This distinction can be functionally characterized: the observed variability in repeated measurements is treated as being due to random causes, whereas systematic causes generate a bias which remains constant across repeated measurements, whose results are then affected in the same way by such causes. The consequence is that, while the assumption of repeatability could be sufficient to assess the presence of random errors, revealing the presence of systematic errors requires performing the measurement of the same measurand by means of different and independently calibrated instruments. Furthermore, the consideration that the effects of these two kinds of causes manifest themselves in statistical versus non-statistical ways led the authors of the GUM to conclude that they "were to be combined in their own way and were to be reported separately (or when a single number was required, combined in some specified way)" (JCGM, 2008: E.1.3).

On this basis a conceptualization was developed that considers the value generated by "the best possible instrument" to be an intrinsic feature of the measurand, traditionally called the *true value* of the measurand and defined as "the value which characterizes a quantity perfectly defined, in the conditions which exist when that quantity is considered" (ISO, 1984: 1.18). The designation "Error Approach" has this origin: due to its experimental component, measurement is unavoidably affected by errors, understood as the difference between the measured value and the true value. Under the assumption of the unknowability of true values, but with the aim of maintaining the operational applicability of the framework, this sharp characterization has sometimes been weakened by instead considering "conventional true values" (of course, the very concept of conventional truth is questionable, to say the least) (see, e.g., ISO, 1984: 3.10) or "reference values" (see, e.g., JCGM, 2012: 2.16), where a reference value "can be a true quantity value of a measurand, in which case it is unknown, or a conventional quantity value, in which case it is known" (JCGM, 2012: 5.18, Note 1). The philosophical justification of the claim of the very existence of a true value of an empirical property is complex and controversial, and we do not discuss it here further.

The important point here is the acknowledgment that "every measurement is tainted by imperfectly known errors, so that the significance which one can give to the measurement must take account of this uncertainty" (BIPM, 1993: Foreword). While errors generate uncertainty in measurement, nothing in principle precludes the possibility that uncertainty has other causes as well: this suggests that measurement uncertainty is an encompassing concept, and justifies the current trend of moving away from the Error Approach and towards the Uncertainty Approach witnessed in both the VIM and the GUM.

3.2.3 The Uncertainty Approach

Like the Error Approach, the Uncertainty Approach can be characterized primarily as a framework that provides functional solutions to implement what we have called an informational strategy to cope with the observed variability of measurement results, and secondarily as a conceptualization that can be included as a background justification for the way in which uncertainty is understood and discussed.[10]

The starting point is that, even when the measurement is not repeated, the information available in the context of the measurement may allow the measurer to acknowledge that the obtained results have a limited quality, due in particular to the quality of the measuring instrument and of the available information on the instrument calibration and on the influence properties. Compared to the Error Approach, the focus here is less on the experimental errors themselves and more on the state of partial knowledge of the measurer, who designs and performs the measurement for the explicit purpose of gaining information on the measurand, with the acknowledgment that "complete" information (whatever this may actually mean) cannot be obtained even by the best possible measurement.[11] As related to measurement results, the concept <measurement uncertainty> emphasizes this incompleteness, and the standardization of the methods for identifying sources of uncertainty and formalizing the quantitative evaluation of their contributions and their combination provides an even more solid common ground for measurement (JCGM, 2008: 0.3):

> A worldwide consensus on the evaluation and expression of uncertainty in measurement would permit the significance of a vast spectrum of measurement results in science, engineering, commerce, industry, and regulation to be readily understood and properly interpreted. In this era of the global marketplace, it is imperative that the method for evaluating and expressing uncertainty be uniform throughout the world so that measurements performed in different countries can be easily compared.

Such an information-oriented standpoint is the basis for a recommendation issued in 1980 by a working group promoted by the International Bureau of Weights and Measures (BIPM) and approved in 1981 by the International Committee of Weights and Measures (CIPM). The traditional classification of kinds of (causes of) error as random or systematic is replaced here with a distinction about the *methods of evaluating* measurement uncertainty (JCGM, 2008: 0.7):

> The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated: A. those which are evaluated by statistical methods, B. those which are evaluated by other means.

This change is the premise for the two key parts of the recommendation.

- First, uncertainties shall be formalized as standard deviations not only when a statistical sample is available, i.e., for "the components in category A", but also in all other cases, i.e.,

[10] The uncertainty that is being addressed in this section (and elsewhere in this and other chapters) is not associated with sampling variability, that is with the uncertainty that is due to a situation where a statistical result is based on a sample from *a population of properties of distinct objects*, where a parameter of the statistical distribution is being estimated. This is usually denoted as *sampling error*, and is primarily more a statistical than a metrological issue.

[11] This may be considered a measurement-specific case of the fundamental distinction between models and modeled entities, sometimes presented in terms of maps versus territory: the only "perfect" map is the territory itself, so that, paradoxically, aiming at a perfect map makes the mapping process useless (the subject of *On exactitude in science*, a delightful short story by Jorge Luis Borges, 1946, 1975). Analogously, a claimed-to-be-perfect measurement would directly exhibit the property under measurement (the perfect representative of itself, indeed), thus making the process of measuring pointless.

for "the components in category B",[12] for which the standard deviation is "based on the degree of belief that an event will occur" (JCGM, 2008: 3.3.5). The list of these components – each then made of a description of the evaluation method and the related standard deviation, called "standard uncertainty" in this context[13] – is included in the *uncertainty budget* (JCGM 2012: 2.33).

- Second, as a consequence of this formalization, the information provided by all uncertainty components in an uncertainty budget can be synthesized in a single outcome: "The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of standard deviations." (JCGM, 2008: 0.7).

Of course, such a position might be considered a pragmatic means of solving the problem of separately reporting statistical and non-statistical components of uncertainty, while simply sidestepping the traditional problem of separately identifying random and systematic causes of errors, as maintained for example by Rabinovich (2005: p. 286).

The subject is complex and widely analyzed in the literature on the science and philosophy of measurement, though for the present purposes we need not discuss it further here.

Box 3.1 – The logic of error / uncertainty propagation

What traditionally has been called *the law of propagation of errors* can be exemplified by the measurement of human body temperature by means of a mercury thermometer. Several possible sources of error can be identified in this measurement, including the finite resolution of the instrument (which is not able to discriminate among temperatures closer to one another than a given threshold), the effect of the temperature and the atmospheric pressure of the environment on the instrument (so that the instrument output changes when the temperature under measurement does not change), and the alteration of the temperature under measurement due to the interaction between the body and the instrument. Under the hypotheses that (1) each of these errors, whether its source is of a statistical nature or not, can be formalized as a standard deviation (thus consistently with the

[12] The terms chosen in the VIM are even more explicit: "Type A evaluation of measurement uncertainty" and "Type B evaluation of measurement uncertainty" (JCGM, 2012: 2.28 and 2.29). The VIM itself provides some examples of Type B evaluations, "based on information (i) associated with authoritative published quantity values, (ii) associated with the quantity value of a certified reference material, (iii) obtained from a calibration certificate, (iv) about drift, (v) obtained from the accuracy class of a verified measuring instrument, (vi) obtained from limits deduced through personal experience" (2.29, Examples).

[13] The term "standard measurement uncertainty" was introduced by the GUM as "uncertainty of the result of a measurement expressed as a standard deviation" (JCGM, 2008: 2.3.1), where the adjective "standard" here plausibly refers to the choice of formalizing all components of measurement uncertainty with the same mathematical tool, i.e., as standard deviations. Whether this is always a sensible position is an open issue, and in any case for less-than-interval properties other tools need to be adopted, for example the interquartile range for ordinal properties and the entropy for nominal properties (Mari et al., 2020). A basic justification of the choice of standard deviations is implicitly given by the GUM itself, which defines <uncertainty (of measurement)> as "parameter [...] that characterizes the dispersion of the values that could reasonably be attributed to the measurand" (2.2.3), thus assuming that, at least in the case of measurement, uncertainty and dispersion can be superposed. That this is generally not the case is clear, as the following quote shows. "Entropy measures the uncertainty associated with a probability distribution over outcomes. It therefore also measures surprise. Entropy differs from variance, which measures the dispersion of a set or distribution of numerical values. *Uncertainty correlates with dispersion, but the two differ*. Distributions with high uncertainty have nontrivial probabilities over many outcomes. Those outcomes need not have numerical values. Distributions with high dispersion take on extreme numerical values. The distinction can be seen in stark relief by comparing a distribution that has maximal entropy with one that has maximal variance. Given outcomes that take values 1 through 8, the distribution that maximizes entropy places equal weight on each outcome. The distribution that maximizes variance takes value 1 with probability 1/2 and value 8 with probability 1/2." (Page, 2018: p. 139, emphasis added). The term "measurement uncertainty" is then taken by the GUM as idiomatic.

CIPM recommendation mentioned above, as then implemented by the GUM), (2) such errors are statistically uncorrelated, and (3) their contribution to the total error is not analytically known, the simplest case of the law of propagation of errors is obtained, which prescribes computing the total error as the square root of the sum of the squares of these standard deviations (i.e., of the variances associated with the errors).

The underlying logic is as follows. For each component X_i it is assumed that a measured value x_i, computed as a sample mean value, and an error, formalized as the standard deviation $s(x_i)$ of the mean, are known. The measurand Y is assumed to be a function of the components, $Y = f(X_1, ..., X_n)$ (in the case of indirect measurement – see Sect. 2.3 and Chap. 7 – f could be the function that computes the measurand Y from the input quantities X_i), so that the measured value y of Y is, as usual, computed as $y = f(x_1, ..., x_n)$. Under the supposition that the errors are sufficiently small and that f is derivable and can be linearly approximated around the n-dimensional point $(x_1, ..., x_n)$, the total error $s(y)$, in turn formalized as a standard deviation, is computed by the first-order approximation of the Taylor series of f, which in the simplest case in which the quantities X_i are not correlated corresponds to

$$s^2(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial X_i} \right)^2 \Bigg|_{X_i = x_i} s^2(x_i)$$

In the case f is not known (hypothesis (3) above), all partial derivatives – each modeling the relative weight of the component X_i to the total error – are assumed to be equal to 1, thus leading to the quadratic sum as in the previous example.

By reinterpreting the errors $s(x_i)$ as standard uncertainties, the GUM has taken this traditional result and assumed that it can be systemically applicable also to uncertainties evaluated by non-statistical (i.e., "type B") methods (for an expanded explanation see the GUM itself, JCGM, 2008, or, in particular, Lira, 2002, and Rossi, 2014).

3.2.4 Basic components of measurement uncertainty

The structure of the measurement process introduced in Sect. 2.2.4 is reflected in a classification of the components of measurement uncertainty.[14] By re-interpreting the abstract structure in Fig. 2.8, still in reference to quantities with unit for the sake of simplicity, some basic components can be identified, as depicted in Fig. 3.3.

[14] This classification is less analytical but possibly more conceptually sound than the list of the "many possible sources of uncertainty in a measurement" proposed in the GUM: "a) incomplete definition of the measurand; b) imperfect realization of the definition of the measurand; c) non-representative sampling – the sample measured may not represent the defined measurand; d) inadequate knowledge of the effects of environmental conditions on the measurement or imperfect measurement of environmental conditions; e) personal bias in reading analogue instruments; f) finite instrument resolution or discrimination threshold; g) inexact values of measurement standards and reference materials; h) inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm; i) approximations and assumptions incorporated in the measurement method and procedure; j) variations in repeated observations of the measurand under apparently identical conditions." (JCGM, 2008: 3.3.2).

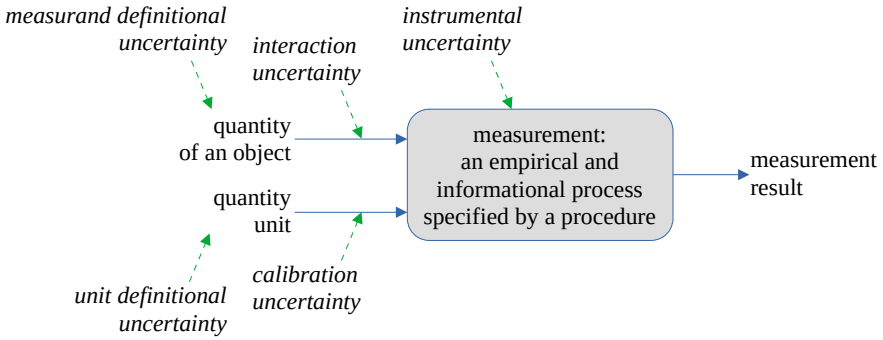


Fig. 3.3 The basic components of measurement uncertainty as related to the abstract structure of measurement (in the case of quantities)

- *Measurand definitional uncertainty.* Measurement is an evaluation of a property of an object, and therefore of an individual property, which is an instance of a general property. Hence, in order to design a measurement process some knowledge of the general property is usually presupposed, and on its basis a model is adopted for the measurand, which in principle should be defined to guarantee that the obtained Basic Evaluation Equation

  measurand = measured value of a property

  conveys meaningful information. This is by no means a trivial assumption: in defining such a model, it may be admitted that the measurand is not completely identified and characterized, thus acknowledging the presence of a non-null measurand definitional uncertainty, "resulting from the finite amount of detail in the definition of a measurand" (JCGM, 2012: 2.27) and therefore being "the practical minimum measurement uncertainty achievable in any measurement of a given measurand" (JCGM, 2012: 2.27 Note 1).[15] A classic example of definitional uncertainty from the physical sciences comes from the measurement of temperature: by defining the measurand as the temperature of a given body, the body itself is (implicitly) modeled as thermally homogeneous, and actual differences of temperature between parts of the body contribute to definitional uncertainty (see also Box 2.3). The human sciences, arguably, regularly contend with many forms of definitional uncertainty. In the case of reading comprehension ability, for example, (a) there might be a lack of clarity regarding the definition of the object that bears the property of reading comprehension ability, insofar as it could be (and, historically, most commonly is) a single human being in isolation, or, in the context of contemporary sociocultural theories (e.g., Mislevy, 2018), it might be a group of people or a single person with a given set of sociocultural resources, which could have practical implications for issues such as whether the examinee should be allowed access to the Internet as they read and attempt to comprehend a text; (b) there might be a lack of clarity regarding the definition of reading comprehension ability in general (e.g., involving questions such as: does it include reading speed? are orthographic fluency and morphemic awareness subcomponents of reading comprehension ability, or causes of it? does reading comprehension ability include metacognitive abilities like critical analysis of textual information, or just "direct" comprehension? and so forth); (c) there might be a lack of clarity regarding how the

[15] There is an unresolved ambiguity here: is definitional uncertainty a component of measurement uncertainty, and thus in fact a standard uncertainty, to be combined with the other components, or the lower bound of the result of such a combination? The GUM is quite clear on this matter by considering definitional uncertainty (the GUM calls it "intrinsic") to be "the minimum uncertainty with which a measurand can be determined, and every measurement that achieves such an uncertainty may be viewed as the best possible measurement of the measurand. To obtain a value of the quantity in question having a smaller uncertainty requires that the measurand be more completely defined." (JCGM, 2008: D.3.4).

general property of reading comprehension ability is instantiated in a given human being, if, for example, an individual is blind and uses a text-to-speech program (is this still considered "reading"?), and so forth. We return to a discussion of special issues faced in the human sciences in Chap. 8. The quantification of measurand definitional uncertainty is an issue that depends on the specific context: we are not aware of any generally applicable technique on this matter.

- *Unit definitional uncertainty.* Measurement is performed as a (direct or indirect) comparison of the quantity of an object with a unit (the more general case of non-quantitative properties is discussed in Chap. 6). Complementary to measurand definitional uncertainty, the quantity unit also needs to be defined before it can enter a comparison process, and nothing in principle prevents such a definition from involving *unit definitional uncertainty*. In the 2019 edition of the International System of units (SI) the definitional strategy is adopted of deducing the definition of the units from the numerical values assumed of some constant quantities, where "the numerical values of the seven defining constants have no uncertainty" (BIPM, 2019: p. 128; see also Sect. 6.3.4): in this case, unit definitional uncertainty is zero.
- *Calibration uncertainty.* The relation between the unit, as defined, and the value selected for the measurand is guaranteed by the calibration of the measuring instrument, which generally cannot be assumed to be perfectly stable, resulting in a non-null *calibration uncertainty*. For example, in measuring the temperature of a body the thermometer could have been calibrated long before the measurement, and in the period since the calibration its structure might have been changed, so that the obtained temperature no longer corresponds to the correct one. In the case of reading comprehension ability, if different texts were being used for different students, and these texts had different reading difficulties, then any process that ignored these differences would be an instance of calibration uncertainty.
- *Instrumental uncertainty.* Again in reference to the Basic Evaluation Equation

  measurand = measured value of a property

  referring the obtained value to the measurand is justified by the quality of the measuring instrument, which is expected to generate an output that is stable in the case of repeated interactions with the object under measurement, and specifically depends on the measurand and not on other properties, i.e., the influence properties. The fact that in this sense the measuring instrument has a limited quality is acknowledged in terms of a non-null *instrumental uncertainty* (JCGM, 2012: 4:24), which is inversely related to the instrument's accuracy (see Sect. 3.2.1). For example, the thermometer used to measure the temperature of a body could be sensitive also to the temperature of the environment, and therefore could produce an indication affected by instrumental uncertainty due to its dependence on properties other than the measurand. In the case of reading comprehension ability, if different students were asked questions by different judges, and these judges expressed the same questions in different ways, this would be an example of instrumental uncertainty.
- *Interaction uncertainty.* Finally, the interaction between the object under measurement and the measuring instrument can alter the state of the object itself. This may occur when acquiring information on physical properties – the so-called "loading effect" – and it is even more usual for psychosocial properties, as for example in most cases of interviews, in which a respondent may be prompted by interview questions to consider issues in a new way. This is acknowledged in terms of a non-null *interaction uncertainty*. In the case of temperature measurement, a sufficiently small body might change its temperature due to its interaction

with an initially colder or warmer thermometer, thus corresponding to an interaction uncertainty. In the case of educational testing, examinees who are asked to respond to a given set of test questions arranged from most to least difficult might conceivably perform worse, on average, than examinees asked to respond to the exact same set of questions arranged from least to most difficult, if their confidence is affected by their experience with the first few questions. Another well-known example in human science measurement relates to "stereotype threat", where people from different sociocultural groups, who may have different assumptions regarding the overall likelihood of success of individuals from their own group on the instrument, tend to respond in ways that are sensitive to those beliefs, especially if their identity as members of the relevant group is made psychologically salient (see, e.g., Steele & Aronson, 1995).

While this classification offers a rich, multidimensional perspective on measurement uncertainty, the aim of providing an overall indication of the quality of the information produced by the measurement requires such components eventually to be combined.

3.2.5 Measurement uncertainty and measurement results

As construed in the Uncertainty Approach and specified by the GUM, measurement uncertainty is a quantity associated with measurement results and inversely related to the quality of the information they convey: the greater the uncertainty, the lower the quality.[16] There is an open debate about what, specifically, is uncertain when measurement uncertainty is stated (the measured value? the measurement result? the estimate of the true value of the measurand? the trustworthiness of the acquired information? ...: see, e.g., the mention in JCGM, 2008: 2.2.4), but the general agreement seems to be that measurement uncertainty is an encompassing entity aimed at summarizing the quality (and quantity) of information acquired through the measurement. The components discussed above summarize the quality-related aspects of a measurement system and, independently of the way they are evaluated, by either statistical or non-statistical methods, they can be in turn summarized into a single, overall *combined standard measurement uncertainty* (JCGM, 2012: 2.31).

The model proposed by the GUM on this matter can be first considered as a black box. By quoting again the BIPM/CIPM recommendation of 1980, "the combined uncertainty and its components should be expressed in the form of standard deviations" (JCGM, 2008: 0.7): from a list of standard deviations, one for each identified component, a standard deviation must be computed as result. There is nothing new in this problem, and in fact the recommendation states that "the combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances" (JCGM, 2008: 0.7). This reinterprets, in the context of the Uncertainty Approach, what is traditionally called the "law of error propagation" (see, e.g., Bevington, 1969: p. 58; see also Box 3.1), which is based on a partial sum of the Taylor series expansion of the function by which a value of the measurand is computed, about the measured value and usually computed only in its first-order terms under the hypothesis of sufficient linearity of the function at the measured value.

The conclusion reached in Chap. 2 about how to report the information obtained by a measurement may be revised accordingly, and then written

[16] Measurement uncertainty is dependent on the quality of measurement results given the available information, not in any "absolute" sense. As remarked by Ignazio Lira, "at first sight this is intuitively correct: if two results of the same quantity are available, the one having a smaller uncertainty will be better than the other. However, by itself the uncertainty says nothing about the *care* put into modelling the measurand, performing the actual measurements and processing the information thus obtained. For example, a small uncertainty may be due to the fact that some important systematic effect was overlooked. Hence, the quality of a measurement can be judged on the basis of its stated uncertainty solely if one is sure that every effort has been taken to evaluate it correctly." (2002: p. 44).

measurand = (measured value of a property, combined measurement uncertainty)

where, if more analytical information were required, the whole uncertainty budget could also be reported.

This relation (or at least its right-hand side term) is to be considered the measurement result, contrary to the tradition still witnessed in the definition of <measurement result> given in the second edition of the VIM: "value attributed to a measurand, obtained by measurement" (BIPM, 1993: 3.1). In other words, from this perspective the measurement uncertainty is assumed to be a constitutive component of the measurement result, and not just an additional, complementary entity. Indeed, "when a measurement result of a quantity is reported, the estimated value of the measurand [...] and the uncertainty associated with that value, are necessary" (BIPM, 2019: p. 127). In the clear words of Lira (2002: p. 43),

> we will […] refrain from referring to the "uncertainty *of* a measurement result". There are two reasons for this. First, a measurement result should be understood as comprising *both* the estimated value of the measurand and its associated uncertainty. Second, once the estimated value is obtained, there is nothing uncertain about it. […] Hence, expressions such as the uncertainty *in knowing the measurand* or the uncertainty *associated with an estimated value* are more appropriate, even if longer sentences result.

This paves the way for extending the very concept of measurement uncertainty to the evaluation of quantities for which the expected value and the standard deviation of the underlying distribution are not sufficiently representative. An example would be where the distribution is strongly asymmetric. More generally, this could encompass ordinal or nominal properties (Mari et al., 2020), for which standard deviations are not meaningful. One solution is to acknowledge that entire probability distributions of values could be reported to convey the available information, on each uncertainty component and then the measurand, as in[17]

measurand = distribution of values of a property

possibly according to the modeling mentioned in Box 3.2. Attributing to the measurand a single value or a distribution of values may in fact be considered the two extreme options, where other strategies are possible for reporting the information acquired by the measurement, so as to convey more information than a single value[18] but less information than an entire distribution. In particular, a measurement result could be reported as a subset of values (usually an interval of values, in the case that the measurand is a quantity), where for discrete cases the greater the number of the values in the subset, the greater the measurement uncertainty, or as a subset of values and a confidence level, i.e., the probability attributed to the subset. This multiplicity of strategies also reflects the variety of tasks involving measurement results: while single values are the usual choice for uncertainty propagation

[17] As an example, let us consider the task to determine the character written in a given ink pattern, called "optical character recognition" (OCR) in the context of Information Technology. If the recognition of a given character from a given pattern is not certain, more than one character could be attributed to the pattern, and in the most general case the result of the recognition is a probability distribution over the chosen alphabet (Mari et al., 2020). Hence, in this case it is a list of distributions, and not of standard uncertainties, that has to be propagated. Due to the analytical complexity of the problem, the GUM framework includes a numerical procedure for such a propagation of distributions, based on a Monte Carlo method (JCGM, 2008b).

[18] For measurands that are quantities, the value is a number that multiplies a unit. In this case the number may actually convey some information about the intended quality of the result through its number of significant digits, so that, for example, "1.23" can be interpreted as including all numbers in the range (1.2250…, 1.2349…). This offers a justification for the admission that "the measurement result may be expressed as a single measured quantity value. In many fields, this is the common way of expressing a measurement result." (JCGM, 2012: 2.9, Note 2). Of course, this is less informative than the standard deviation format, except if it is also assumed that the distribution in the range is of a particular kind, such as a uniform distribution.

and computing functions in indirect measurement, and of course for daily, non-scientific uses, intervals of values may be more suitable in decision-making applications, for example conformity assessment or when investigating the compatibility of two measurement results.

Box 3.2 – Another perspective on (un)certainty

Given that, particularly in the context of Bayesian interpretations, probability is considered to be the logic of uncertainty,[19] one could wonder why the default quantitative model of measurement uncertainty is standard deviation, instead of probability itself. The difference is not trivial: while a probability is a pure number, a standard deviation has the same dimension as the measurand.

There is, first, a plausible historical reason for this: measurement uncertainty is the offspring of measurement error, which is indeed the difference from the true value, with which it then shares its dimension. But another reason seems not less important. Reporting a measurement result in terms of a single measured value, together with a standard uncertainty whenever appropriate, is a convenient choice given that most mathematical models (e.g., physical laws) are designed to operate on values, not on subsets / intervals of values or more complex entities like probability distributions over values (the numerical propagation of distributions through analytical functions has only recently become feasible thanks to the availability of efficient computational tools, as presented in JCGM, 2008b). But if the information on the value of the measurand is reported as a single value, what remains for providing information about the quality of the result is some index of the expected dispersion of the measured value, thus under the questionable assumption – already discussed in Footnote 13 – that dispersion and uncertainty are basically the same concept.

However, by relaxing the condition of single measured value a more general and expressive modeling framework can be adopted, as follows. Let us suppose that the information empirically acquired on the measurand is summarized by means of a probability mass or density function. From such a function several coverage subsets / intervals can be obtained, each of them with an associated coverage probability (as a well-known example, for a Gaussian function a coverage interval centered in the expected value and whose half width is two standard deviations corresponds to a coverage probability of about 0.95). The quality of a measurement results has then two dimensions: the width of the coverage interval is about the inverse of the *specificity* of the reported information, while the coverage probability is about the *certainty* attributed to the interval. As it is reasonable, once the information on the measurand has been empirically acquired, and therefore once the underlying probability distribution is chosen, reporting a more specific information makes it less certain, and vice versa. This provides us with some added flexibility in reporting measurement results, by balancing their specificity and certainty: for example, if the length of a rod can be reported in centimetres, the result will be more certain than if reported in millimetres.

Of course, there is a sharp difference between this concept of certainty and what is today commonly considered measurement uncertainty: in this framework, certainty is a probability, thus ranging from 0 (deemed to be impossible) to 1 (deemed to be certain).

[19] For example, according to Bruno De Finetti, "the assessing of probabilities, as expressions of subjective feeling, and reasoning with probabilities, obeying objectively necessary conditions of coherency, constitute the logic of uncertainty" (1972: p. xiii).

3.3 The operational context

Measurement is a process designed and performed in a context that is in fact structurally more complex than the one introduced in Chap. 2 and depicted in Fig. 2.8, for at least the following reasons:

- the quantity unit is defined independently of the specific measurement problem, and is made available through a *metrological system*, and
- the comparison between the measurand and the unit, and therefore the obtained measured value, is generally affected by other properties, which reveal the presence of a *measurement environment*.

Through the consideration of these contextual elements, as depicted in Fig. 3.4, let us switch from an abstract and conceptual interpretation of measurement to one that is more concrete and operational. We discuss here the case of quantities and defer the treatment of non-quantitative properties and their values to Chap. 6.

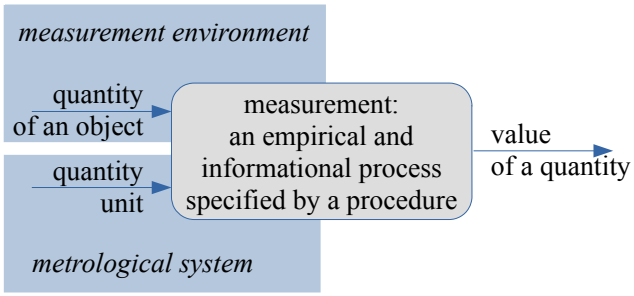


Fig. 3.4 The broad context of measurement (in the case of quantities)

3.3.1 The metrological system

Measurement is a process that enables the quantity-related comparison of objects through a process of delegation: for any two objects a and b both having a general quantity Q (say, length or reading comprehension ability), the information that a and b are empirically indistinguishable with respect to Q, $Q[a] \approx Q[b]$, can be obtained not only through their direct comparison (e.g., by the comparison of the extreme points of two rods, possibly mediated by a third rod, to evaluate their lengths, or by the synchronous comparison of two individual readers by a judge, to evaluate their reading comprehension abilities) but also by means of the independent measurement of the two quantities and the comparison of the obtained values. If the measured value of the lengths of two rods is the same, then the two rods are inferred to have the same length; if the measured value of the reading comprehension abilities of two individuals is the same, then the two individuals are inferred to have the same reading comprehension ability.

Hence, through measurement values operate as mediators for the comparison of quantities of objects. The meaning of these equalities is that the chosen unit q_{ref} is a quantity of the same kind as $Q[a]$ and $Q[b]$, and $Q[a]$ and $Q[b]$ have the same relation with q_{ref}, in the sense that if $Q[a] = n_1\ q_{ref}$ and $Q[b] = n_2\ q_{ref}$ then $n_1 = n_2$. In principle, this requires the unit q_{ref} to be accessible for its comparison with the measurands, $Q[a]$, $Q[b]$, ..., even if the measurements are performed in different places and times: the widespread availability of the unit needs then to be somehow guaranteed.

In some cases, the only practical solution is to produce and disseminate multiple objects that realize the definition of the unit. In the tradition of physical measurement, the organizational and technical structure aimed at guaranteeing this is called a *metrological system*. Whenever it is possible to infer the information on the comparison to the unit from the comparison with the quantity realized by a

replicated object, a measurement result is said to be *metrologically traceable* (JCGM, 2012: 2.41) to the unit. Hence, in order for one to be able to make the inference that $Q[a] \approx Q[b]$ from $Q[a] = n\ q_{ref}$ and $Q[b] = n\ q_{ref}$ even if the two measurements were performed in different places and times, the metrological traceability of the two results to the same unit must be guaranteed by an effective metrological system.

The quality of metrological systems is traditionally maximized through a structural strategy of hierarchical delegation: the definition of the unit is first realized in a *primary measurement standard* (JCGM, 2012: 5.4), which is then replicated in some secondary standards that are disseminated, which in turn are replicated and disseminated, and so on, thus generating *traceability chains* (JCGM, 2012: 2.42) of standards, under the responsibility of National Metrology Institutes and accredited calibration laboratories. Mari and Sartori (2007) show that under given conditions this strategy is both efficient and effective. Indeed, a metrological system, as a network of measurement standards and measuring instruments connected through calibrations,

- is connected by a relatively small number of calibrations, each corresponding to an edge of the network, and therefore *the system is efficient* because its global costs are relatively small,
- and the average length of the traceability chains (corresponding to the average shortest path length between nodes) is also small, and therefore *the system is effective* because each calibration reduces the quality of the provided information, so that minimizing the length of a traceability chain maximizes the quality of the measurement results.

Hence, metrological systems reproduce the structure of small-world networks, which at the same time are connected but guarantee a small number of degrees of separation (Watts & Strogatz, 1998).

In the case of reading comprehension ability (like most other psychosocial properties), there are no direct reference objects, and therefore such a system of reference properties and calibrated standards does not pertain. However, responses to test questions are informational, and hence, if they are seen as stand-in reference objects, then a similar system can be constructed, using comparisons between the scored responses to the questions in the original sample, and a second sample of responses (Wilson, 2013). Just as in the discussion above, the length of the traceability chain can be kept relatively short by always using the original sample to calibrate new sample sets. A logic built in this way can be seen as functionally analogous to a metrological system (Maul et al., 2019).

3.3.2 The measurement environment

As presented in Chap. 2, a measurement result reported as the Basic Evaluation Equation

$$Q[a] = x\ q_{ref}$$

(again neglecting measurement uncertainty) assumes that the (direct or indirect) comparison between the measurand $Q[a]$ and the unit q_{ref} depends only on these two quantities. However, in real-world contexts the comparison may be affected by influence properties, of the object under measurement or the environment, including the measuring instrument itself: if an influence property changes, the measurement result could also change even if the measurand remains the same.

For example, returning to an example discussed above, the length of a rigid body could be measured by a caliper whose structure is sensitive to the environmental temperature in such a way that different indications are obtained for the same measurand when the temperature changes: in this case, then, environmental temperature is an influence property. In the case of the measurement of reading comprehension ability, an influence property might be the specific content of the passages a person reads: a person with strong prior knowledge of the content may be more likely to successfully answer

related questions than someone less familiar with the content, even though they may have the same reading comprehension ability. This shows that the information reported by a Basic Evaluation Equation is in fact affected by the context in which the measurement is performed.

There are two complementary structural strategies for taking into account the presence of the influence properties.

One strategy aims at making the measurement less and less sensitive to influence properties, by improving the measuring instrument and the overall measurement procedure. In the case of length measurement, the process could be performed in a carefully temperature-controlled environment. In the case of a test of reading comprehension ability, an example of an attempt to reduce the effects of an influence property would be to deliberately write passages about a topic known to be equally unfamiliar to all examinees. The experimental abilities of measurers have to do with designing and implementing strategies for this purpose.

The other strategy exploits the fact that the measurand is the property *intended* to be measured, and therefore that must be defined, explicitly or implicitly. As exemplified in Box 2.3, the measurand could be then defined by including the specification of some environmental conditions in the definition itself, thus changing the role of the corresponding properties, from influence properties to components of the measurand, or vice versa by broadening the definition so as to make it explicitly indifferent to some influence properties. Chap. 7 includes a discussion about the complex subject of measurand definition. For example, it might be specified that the length of a rigid body is considered at a given temperature, or that reading comprehension ability simply pertains to one's ability to comprehend a given set of texts regardless of whether this comprehension is based solely on semantic processing of the texts or is also aided by prior knowledge of the content of the texts, in which case prior content familiarity is seen as a component of reading comprehension ability rather than an influence property.

3.4 The conceptual context

For performing a measurement, in addition to the operational conditions discussed in the previous section, some conceptual conditions also need to be fulfilled: an existing property of an object has to be identified as the property intended to be measured, leading to a *model of the measurand*, and a suitable process of property evaluation has to be designed, leading to a *model of the measurement*. This last section of the chapter is devoted to a preliminary discussion of these two aspects of the measurement problem, thus setting the strategic context for the more careful and extensive analysis developed in the chapters that follow.

3.4.1 Measurement and property identification

Performing a measurement presupposes that something is there to be measured: as previously mentioned, this is a property of an object. In most cases of mature measurement practice, both the general property and the object are well-known and clearly identified before and independently of the measurement. Given the well-developed status of physics, this is the usual case for measurement of physical properties: it is physics itself that guarantees that general properties such as length, mass, and time duration are sufficiently well-defined, and in fact inter-defined, in a network of general properties (and more specifically quantities) connected by equations globally known as the *International System of Quantities* (ISQ) (see ISO, 2009: 3.6 and JCGM, 2012: 1.6). In this context a measurement problem starts from a previously defined general property and only requires that one identifies the individual

property intended to be measured as an instance of that general property.[20] Of course, as physicists discover new properties, and seek to measure them, it may be that at least initially these assumptions cannot be met.

Thus, things are not always so simple. In the case of physical quantities, interesting examples have been studied of situations in which measurements were instrumental in the very identification / definition of the general property, a well-known case being temperature, as analyzed in particular by Hasok Chang (2007). In these cases the neat sequential procedure – from the assumption of an already-defined general property and a preexisting measuring instrument, to the identification of an instance of that property as the measurand and then the design and operation of a measuring instrument – becomes a complex loop in which the distinctions between the activities of defining the property, constructing the measurement system, and performing the measurement are blurred, and one might operate by measuring without a clear idea of what one is measuring. It may happen, as in the words of Thomas S. Kuhn, that "many of the early experiments involving [a new instrument] read like investigations of that new instrument rather than like investigations with it" (1961: p. 188).[21]

In the context of the human sciences, which currently lack anything like an ISQ, this situation of general property definition intertwined with measurement is not unusual. New variables may be readily obtained via computation, and without a system such as the ISQ to establish that properties are well-defined, such variables are not necessarily the formal counterpart of empirical properties. It is indeed not hard to provide examples of variables, such as the "hage" of a person obtained as the product of her height and age (Ellis, 1968: p. 31), which can be computed very accurately and fulfilling all expected requirements about uncertainty propagation etc., and nevertheless do not seem to correspond to any property of individual humans. Less trivially, this problem becomes critical in the context of complex concepts such as the social status of an individual, the quality of the research performed by a team, the performance of a company, or the wealth of a nation, all of which are sometimes claimed to be measurable by computing given mathematical functions from empirically collected data. In these cases one can interpret measurement as a tool not only for the acquisition of information on the measurand, but also, and even before, for gaining knowledge about the general property under consideration.

The case of reading comprehension ability is interesting in this respect, given how it has changed historically. In the 19th century, one procedure for checking whether students could read was that they were asked to read a text aloud, and often also asked to recite parts of the text from memory (Matthews, 1996; Smith, 1986). Thus, at that point of time, the (implicit) property definition was something like "ability to accurately read a text out loud". With time, it was realized that students could succeed at such tasks without understanding the content of the text passage. This led to the advent of silent reading tests, where reading comprehension ability was assayed by the interaction of

[20] In the context of metrology it is usual to use the expression "measurand definition" (from which, e.g., "definitional uncertainty", JCGM, 2012: 2.27). Under the assumption that properties of objects are empirical, strictly speaking what can be defined is not a measurand but the concept of it (consider the parallel case of objects: what can be defined is not, for example, a rod, but the concept of a rod): a measurand can be instead identified, through a sufficiently specific definition or, more simply but less usefully, a direct reference ("the measurand is the length of that rod" uttered while indicating a rod). Some strategies of measurand identification / definition are discussed in Box 2.3.

[21] The case of temperature is again exemplary of the problems that can be encountered. According to Chang (2007: p. 4): "How do we know that our thermometers tell us the temperature correctly, especially when they disagree with each other? How can we test whether the fluid in our thermometer expands regularly with increasing temperature, without a circular reliance on the temperature readings provided by the thermometer itself? How did people without thermometers learn that water boiled or ice always melted at the same temperature, so that those phenomena could be used as 'fixed points' for calibrating thermometers? In the extremes of hot and cold where all known thermometers broke down materially, how were new standards of temperature established and verified? And were there any reliable theories to support the thermometric practices, and if so, how was it possible to test those theories empirically, in the absence of thermometry that was already well established?".

the reader with comprehension questions (Pearson, 2000), generically called "items". Thus, at this later point, the property definition (still implicitly) changed to something like "ability to demonstrate understanding of the content of the text (and the question)". An early test of this sort was developed by Frederick Kelly (1916), and an example question from which is shown in Fig. 1.1. The questions were chosen to be likely to generate incontrovertibly either correct or incorrect responses. The indication of a student's reading comprehension was then the number, or percentage of, the test items the student answered correctly.

3.4.2 Measurement and measure

In the framework of necessary conditions for measurement introduced in Sects. 2.2 and 2.3 – measurement as an empirical and informational process, designed on purpose, whose input is an empirical property of an object and that produces information in the form of values of that property – a further specification is appropriate here, about the very distinction between measurement and measure.

The ancient Greek verb for <to measure> has the root "metr-" (μετρ-), from which the term "metrology" derives, highlighting the relation of the concepts <to measure> and <measurement>. However, the relation between the concepts designated by the *nouns* "measure" and "measurement" is more delicate.

The Euclidean tradition has been described as "the earliest contribution to the philosophy of measurement available in the historical record" (Michell, 2005: p. 288)[22], as witnessed by the oft-quoted first definition of Book 5 of the Euclid's *Elements*: "A magnitude is a part of a[nother] magnitude, the lesser of the greater, when it *measures* the greater" (Euclid, 2008: V.1, emphasis added). Hence, the hypothesis that the noun "measurement" and the verb "to measure" refer to the same entity appears plausible, so that a theory of measurement and a theory of measure would be more or less the same thing, or at least they would be inherently related. This position is further evidenced, for example, in the claim that "to understand measurement theory it is necessary to revisit the theory of integration and, in particular, Lebesgue measure theory" (Sawyer et al., 2013: p. 90).

However, suspicions about the equivalence of <measurement> and <measure> might arise from a sufficiently careful reading of Euclid's work itself, which is not really about measurement as we intend it. For example, in the introduction to an English translation of the Elements one can read that "in the geometrical constructions employed in the Elements [...] *empirical proofs by means of measurement are strictly forbidden*" (Euclid, 2008: introductory notes, emphasis added). Let us indeed compare the above-mentioned definition by Euclid, with the following one, again from the Elements but now taken from Book 7: "a number is part of a(nother) number, the lesser of the greater, when it *measures* the greater" (VII.3, emphasis added). While the two quoted sentences refer to different entities (magnitudes, μεγέθη, and numbers, ἀριθμοὶ), in both the relation is said that one entity *measures* (καταμετρῇ) the other. Hence, as derived from the Euclidean tradition, "to measure" does not necessarily have an empirical meaning, and in fact the Euclidean <to measure> is coextensive with <to be (an integer) part of> (Mari et al., 2017). Consistently with this position, then, a "measure of a number is any number that divides it, without leaving a reminder. So, 2 is a measure of 4, of 8, or of any even number; and 3 is a measure of 6, or of 9, or of 12, etc." (Hutton, 1795). The conclusion is that "measure" and "to measure" have (at least) two distinct meanings: one is empirical, and is indeed related to measurement, and the other is mathematical; this twofoldness, already recognized by Bunge (1973), has often been confused.[23]

[22] More broadly, the importance of this contribution is also highlighted by the consideration that "Euclid's *Elements*, written about 300 BC, has probably been the most influential work in the history of science" (Suppes, 2002: p. 10).

Perhaps unsurprisingly, on this conceptual basis a "measure theory" has developed, where "a measure is an extended real valued, non-negative, and countably additive set function μ, defined on a ring R, and such that μ(0) = 0" (Halmos, 1950: p. 30): that is, it is a mathematical entity. Whether and how measure theory is related to a theory of measurement, and more generally to measurement science, is an issue that we discuss in Chap. 6, in the section about the measurability of non-quantitative, and specifically non-additive, properties. But it should be clear now that "measure" is not just a synonym of "measurement" and, most importantly, that "quantification" is not just a synonym of "measurement": not every quantification is a measurement, and it could even be accepted that non-quantitative properties may also be measured. Thus, one can see the wisdom in the VIM's avoidance of any use of the noun "measure" (except in the idiomatic term "material measure") to reduce ambiguity,[24] and its adoption of "measurement result" to designate the outcome of the process. This is the lexical choice that we make here too.

The operational and conceptual issues discussed in this chapter provide a basis for our analysis of philosophical perspectives on measurement, to which the next chapter is devoted.

References

Altman, D. G. & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal, 308(6943)*, 1552.

Bevington, P. R. (1969). *Data reduction and error analysis for the physical sciences*. New York: McGraw-Hill.

Borges, J. L. (1975). On exactitude in science. In J. L. Borges, *A universal history of infamy*. London: Penguin Books (first published in Spanish, 1946).

Bunge, M. (1973). On confusing 'measure' with 'measurement' in the methodology of behavioral science. In M. Bunge (Ed.), *The Methodological unity of science* (pp. 105–122). Dordrecht: Reidel.

Chang, H. (2007). *Inventing temperature – Measurement and scientific progress*. Oxford: Oxford University Press.

De Finetti, B. (1972). *Probability, induction and statistics -- The art of guessing*. Wiley: London.

Ellis, B. (1968). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

[23] Consider, as a significant case, what Michell and Ernst wrote on this matter:: "there are two sides to measurement theory: one side (emphasized in the modern era) at the interface with experimental science, the other side (emphasized in the classical) at the interface with quantitative theory" (1996: p. 236). But these are two sides of measure, not measurement. This sentence is excerpted from the introduction that Michell wrote to the English translation of the 1901 paper by Hölder on the axioms of quantity. This confusion was worsened by their choice of translating in the title of Hölder's paper the German noun "mass" as "measurement" rather than "measure": they used "The axioms of quantity and the theory of measurement", instead of "The axioms of quantity and the theory of measure" (Mari et al., 2017).

[24] In reference to the black-box model we have just discussed, the noun "measure" is sometimes used to refer to each of the three elements of the model: the input property, the process, the output value. Just as an example, Isaac Newton famously began his Principia with the following two definitions: "Definition I. The Quantity of Matter is the measure of the same, arising from its density and bulk conjunctly. Definition II. The Quantity of Motion is the measure of the same, arising from the velocity and quantity of matter conjunctly." (1724: p. 1). Here the concepts <quantity> and <measure> appear to be equivalent. Very interesting on this matter is the following quote from Leonhard Euler: "Whatever is capable of increase or diminution is called *magnitude*, or *quantity*. [...] Mathematics, in general, is the science of quantity; or, the science which investigates the means of measuring quantity. [...] Now, we cannot measure or determine any quantity, except by considering some other quantity of the same kind as known, and pointing out their mutual relation. [...] So that the determination, or the measure of magnitude of all kinds, is reduced to this: fix at pleasure upon any one known magnitude of the same species with that which is to be determined, and consider it as the measure or unit; then, determine the proportion of the proposed magnitude to this known measure. This proportion is always expressed by numbers; so that a number is nothing but the proportion of one magnitude to another arbitrarily assumed as the unit." (1765: p. 1-2).

Euclid's Elements of geometry, the Greek text of J. L. Heiberg (1883-1885) edited, and provided with a modern English translation, by Richard Fitzpatrick (2008). Retrieved from farside.ph.utexas.edu/Books/Euclid/Euclid.html

Euler, L. (1765). *Elements of algebra.* Translated into English by J. Hewlett, London.

Ferrero, A., & Salicone, S. (2006). Fully comprehensive mathematical approach to the expression of uncertainty in measurement. *IEEE Transactions on Instrumentation and Measurement, 55(3)*, 706–712.

Giordani, A., & Mari, L. (2014). Modeling measurement: error and uncertainty. In M. Boumans, G. Hon, & A. C. Petersen (Eds.), *Error and uncertainty in scientific practice* (pp. 79–86). Pickering & Chatto, 2014.

Hacking, I. (1975). *The emergence of probability – A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge: Cambridge University Press.

Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.

Halmos, P. R. (1950). *Measure theory*. Litton Educational Publishing (reprinted by Springer-Verlag, 1974).

Hutton, C. (1795). *A mathematical and philosophical dictionary*. London: Johnson.

Idrac, J. (1960). *Measure et instrument de mesure*. Paris: Dunod.

International Bureau of Weights and Measures (BIPM) (2019). *The International System of Units (SI) ("SI Brochure")* (9th ed). Sèvres: BIPM.

International Bureau of Weights and Measures (BIPM) and other six International Organizations (1993). *International Vocabulary of Basic and General Terms in Metrology (VIM)* (2nd ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML), International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP).

International Organization for Standardization (ISO) and other three International Organizations (1984). *International vocabulary of basic and general terms in metrology (VIM)* (1st ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML).

International Organization for Standardization (1994). *ISO 5725-1:1994, Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions*. Geneva: ISO.

International Organization for Standardization (2022). *ISO 80000-1:2022, Quantities and units – Part 1: General* (2nd ed.). Geneva: ISO.

Joint Committee for Guides in Metrology (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2008). *JCGM 101:2008, Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.; 2008 version with

minor corrections). Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Kelly, E. J. (1916). The Kansas silent reading tests. *Journal of Educational Psychology, 7*, 63–80.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52(2)*, 161–193.

Lira, I., (2002). *Evaluating the measurement uncertainty: Fundamentals and practical guidance*. London: Institute of Physics Publishing.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Mari, L., Narduzzi, C., Nordin, G., & Trapmann, S. (2020). Foundations of uncertainty in evaluation of nominal properties. *Measurement, 152*, 1–7.

Mari, L., & Sartori, S. (2007). A Relational Theory of Measurement: traceability as a solution to the non-transitivity of measurement results. *Measurement, 40*, 233–242.

Matthews, M. (1996). *Teaching to read*. Chicago, IL: University of Chicago Press.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770.

Mencattini, A., & Mari, L. (2015). A conceptual framework for concept definition in measurement: the case of 'sensitivity'. *Measurement, 72*, 77–87.

Michell, J. (2005). The logic of measurement: a realist overview. *Measurement, 38*, 285–294.

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement. Translated from Part I of Otto Hölder's German text "Die axiome der quantität und die lehre vom mass", *Journal of Mathematical Psychology. 40(3)*, 235–252.

Mislevy, R. J. (2018). Sociocognitive foundations of educational measurement. Milton Park: Routledge.

Newton, I. (1724). *The mathematical principles of natural philosophy.* Translated into English by A. Motte, London.

Page, S. E. (2018). *The model thinker*. New York: Basic Books.

Pearson, P. D. (2000). Reading in the 20th century, In T. Good (Ed.), *American education: Yesterday, today, and tomorrow. Yearbook of the National Society for the Study of Education* (pp. 152–208). Chicago: University of Chicago Press.

Possolo, A. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST measurement results*. NIST Technical Note 1900. Retrieved from www.nist.gov/publications/simple-guide-evaluating-and-expressing-uncertainty-nist-measurement-results

Rabinovich, S. G. (2005). *Measurement errors and uncertainties – Theory and practice* (3rd ed.). New York: Springer.

Rossi, G. B. (2014). *Measurement and probability – A probabilistic theory of measurement with applications*. Dordrecht: Springer.

Sawyer, K., Sankey, H., & Lombardo, R. (2013). Measurability invariance, continuity and a portfolio representation. *Measurement, 46*, 89–96.

Smith, N. B. (1986). *American reading instruction*. Newark: International Reading Association.

Speitel, K. (1992). Measurement assurance. In G. Salvendy (Ed.), *Handbook of Industrial Engineering* (pp. 2235–2251). New York: Wiley.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology, 62(1)*, 26–37.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford, CA: CSLI Publications.

Tal, E. (2019). Individuating quantities. *Philosophical Studies, 176(4)*, 853–878.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393*, 440–442.

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*, 3766–3774.

Chapter 4.
Philosophical perspectives on measurement

This chapter aims to present a brief conceptual history of philosophical thinking about measurement, concentrating in particular on the issues of objectivity and subjectivity, realism and nonrealism, and the role of models in measurement, as well as a discussion of how these philosophical issues have shaped thinking and discourse about measurement in both the human and physical sciences. First, three perspectives on measurement and its epistemic status are discussed, grouped as (a) naive realism, (b) operationalism, and (c) representationalism. Following this, we discuss how these perspectives have informed thinking about the concept of validity in the human sciences, and how they have influenced the way in which measurement is characterized in different contexts as being dependent on empirical and/or mathematical constraints. We then attempt to synthesize these perspectives and propose a version of *model-dependent realism* which maintains some of the elements of each of these perspectives and at the same time rejects their most radical aspects, by acknowledging the fundamental role of models in measurement but also emphasizing that models are always models *of* something: the empirical components of measurement are designed and operated to guarantee that, via such models, measurement results convey information on the intended property. The analysis also provides a simple explanation of two of the most critical stereotypes that still affect measurement science: the hypotheses that (1) measurement is quantification, which hides the relevance of the empirical component of the process, and that (2) measurement is only a process of transmission and presentation of pre-existing information, usually intended as the "true value" of the measurand, which instead neglects the role of models in the process.

4.1 Introduction

Measurement is often considered an effective method for acquiring information about the empirical world. When claiming to have measured something, the clear implication is that the information one has obtained is of higher quality in some sense (more dependable, more accurate, more objective, ...) than the information acquired via other means, such as conjecture based on personal intuition. However, as previously noted, it is not always clear on what grounds activities referred to as being measurements earn the epistemic and social authority commonly afforded to them: in other words, what is it about an activity being a measurement that secures the quality of its results? And can an answer to this question be given independently of consideration of the particular area of application?

Providing answers to such questions necessitates confronting some of the most fundamental questions humanity has asked regarding what can be known and how we know it, such as: what *is there*, and what is it like? Of what is it possible to have knowledge? How is knowledge acquired? How are knowledge claims justified, or how should they be justified? How is knowledge codified and communicated? Given how central these questions have been in the development of philosophy, it is not surprising that the history of thinking about measurement is deeply intertwined with the history of philosophy in general, and in particular of the branches of philosophy concerned with science, ontology, epistemology, and semiotics. In order to better appreciate contemporary perspectives on

measurement, and in particular the issue of whether it is indeed possible to formulate a conception of measurement applicable across the sciences, this chapter aims to provide a brief background on the history of philosophical perspectives on measurement, and to explore their impact on thinking and discourse about measurement in both the human and physical sciences.

4.1.1 Measurement between objectivity and subjectivity

The increasing interest in the measurement of psychosocial properties, and the accompanying widening of the scope of activities referred to as instances of measurement, has precipitated a new wave of interest in the philosophical foundations of measurement, starting with the basic question of what measurement actually *is*, or, more operationally, what conditions have to be fulfilled for a process to be accepted as a measurement.

We take here as a premise that there is no single, inherent meaning of "measurement" which exists independently of practice and context, and that some amount of convention is unavoidable. The common observation that there are multiple and sometimes incompatible definitions of what measurement is throughout the scientific, technical, and philosophical literatures might be simply interpreted as evidence that measurement is a many-faceted activity, and that this multiplicity is somehow irreducible. However, as already noted, an "anything goes" relativism or conventionalism – which might assert that *any* process of property evaluation could be considered a measurement if it is agreed to be so by a relevant community – would not justify the epistemic authority attributed to measurement.

The issue is sometimes cast in terms of the issue of *objectivity*: in brief, scientific inquiry is generally expected to provide knowledge that is independent of the particular perspective(s) of individuals involved in its acquisition. While a radically subjectivistic (or relativistic) position might claim that any human evaluation is by definition a measurement, the scientific, technological, and social history of measurement has, in contrast, emphasized an objectivist view. For many centuries, indeed, measurement has been primarily understood as applicable only to physical quantities, which are generally believed to exist independently of human perception or thought – i.e., *objectively*. Hence objectivist presuppositions have accompanied the development of scholarship on measurement, concomitant with the presumption that measurement is the ground on which scientific and technical developments are based whenever dependable data is required, a "protocol of truth" in the classical terminology of philosophy (Margenau, 1958). In the 19th century, the possibility of measuring properties of human beings – and, in particular, properties of the mind and of conscious experience, which of course (and indeed, tautologically) do *not* exist independently of human perception or thought – finally became a topic of research and development, first about psychophysical events (i.e., the effects of physical phenomena on human perception), and then about psychological properties such as, notoriously, intelligence. More recently, a wide range of human attributes have become the targets of measurement, including, for example, domains of knowledge and skill, cognitive and physical abilities, aspects of personality, affective and motivational characteristics, psychological conditions, attitudes, values, and preferences. In particular, this raises the issue of whether properties must exist independently of human thought – i.e., be *ontologically* objective, in the terminology of John Searle (e.g., 1992; see also Maul, 2013) – in order to be proper objects of measurement, or scientific inquiry more generally, which is in general expected to produce objective information, i.e., to be *epistemically* objective.

The tension between objectivity and subjectivity in measurement can even be regarded as the basic philosophical problem of measurement (Mari, 2003): *what distinguishes measurement from generic evaluation?*[1]

In the course of history, it has been suggested that measurement can be characterized in reference to:[2]

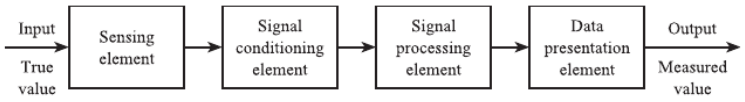
- *ontic* reasons: e.g., measurement is a process designed to discover the values that properties of objects inherently have;
- *epistemic* reasons: e.g., measurement is a process designed to produce true, or at least credible, information on the properties intended to be measured;
- *pragmatic* reasons: e.g., measurement is a process designed to be adequate for its goal of obtaining information on the properties intended to be measured; and/or
- *formal* reasons: e.g., measurement is a process designed to evaluate properties in a consistent way by means of symbols.

These positions are not mutually exclusive. It seems fair to say that any well-conceived standpoint on measurement would acknowledge the validity of more than one (and possibly all) of these positions, though there may be room for differences in terms of the balance of their relative importance. On this basis, the next section is devoted to exploring some of the most influential perspectives on these issues, and in particular to how they each lead to different characterizations of the necessary and sufficient conditions for measurement.

4.2 Characterizing measurement

A wide range of characterizations of measurement are found throughout the scientific and technical literature, witnessing both the wide interest in the subject and its complexity. This multiplicity can be interpreted according to some complementary criteria (Mari, 2013), as in Table 4.1.

Table 4.1 Comparison of definitions of <measurement> according to different criteria

Criterion	**Exemplary definition / characterization**
Is measurement characterized by the structure of the process?	"[Any] measurement system consists of several elements or blocks. It is possible to identify four types of element, although in a given system one type of element may be missing or may occur more than once."  (Bentley, 2005)
Or by the results it produces?	"Measurement is essentially a production process, the product being numbers." (Speitel, 1992)

Criterion	**Exemplary definition / characterization**

[1] As noted in Footnote 21 of Chapter 2, we use the term "evaluation" to refer to an attribution of a value to the property of an object.

[2] For an introduction to the philosophical understanding of measurement, see also Tal (2020).

Does measurement imply comparison to a reference, possibly a unit?	"Measurements are executions of planned actions for a qualitative comparison of a measurement quantity with a unit." (DIN, 1995)
Or not?	"Measurement is the process of empirical, objective assignment of numbers to the attributes of objects and events of the real world, in such a way as to describe them." (Finkelstein, 1994)

Criterion	**Exemplary definition / characterization**
Are numbers required products of measurement?	"Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real, as the case may be." (Russell, 1903)
Or not?	"The only decisive feature of all measurements is symbolic representation; even numbers are in no way the only usable symbols. Measurement permits things (relative to the assumed measuring basis) to be presented conceptually, by means of symbols." (Weyl, 1949)

Criterion	**Exemplary definition / characterization**
Are experimental activities required to perform a measurement?	"Measurement is the set of empirical and informational operations performed by means of suitable devices interacting with the system under measurement with the purpose of assigning a value of a quantity assumed as parameter of the system" (UNI, 1984, translated from Italian)
Or not?	"Measurement is the assignment of numerals to objects or events according to rule, any rule." (Stevens, 1959)

For each criterion, the incompatibility of the two positions is manifest. And while these incompatibilities have not prevented important advancements in measurement science, such a confused and confusing situation is clearly not optimal. We propose that one can understand this multiplicity as being the result of some stereotypes that affect measurement (Mari et al., 2017): each of these positions is grounded on some reasonable basis and many are attractive for their simplicity, but the interpretations they provide each miss some key aspects of the complex concept of measurement. In this section we analyze those which we consider to be the three main stereotypes of measurement, as related to *naive realism*, *operationalism*, and *representationalism*.

4.2.1 Naive realist perspectives on measurement

Realism builds upon the hypothesis that the world exists independently of us and our understanding of it. Realist perspectives may acknowledge the importance of models, but emphasize that any model is always a model *of* something that exists independently of the model. Alternatively, a realist perspective might deny the importance of models altogether, thus with the implication that our knowledge of reality can be *direct*, in the sense of being unmediated by models or unfiltered by forms of conceptual and linguistic schemes; such a view is sometimes referred to as *naive realism*.

In the case of measurement, such a perspective is sometimes associated with the perspective that each property of each object has, inherently, a value – sometimes called "true value" – and that measurement would simply aim at *discovering* such a value.[3] This implies that measurement is a sort of transmission or communication process, which, in the ideal case, perfectly transfers an entity from the object under measurement to the measuring instrument and thus makes it in some sense observable. Hence, this "true value" is "the value that would be obtained by a perfect measurement" (Bell, 1999), with the consequence that, formally, "measurement [...] in a deterministic ideal case results in an identity function" (Rossi, 2006: p. 40). This fits well with an abstract understanding of the entities that constitute the domain of the measurement processes, conceived in analogy with lengths of line segments in an abstract mathematical space and therefore in continuity with the Euclidean standpoint.

Such a view is sometimes traced back to Greek antiquity. Indeed, as described by Aristotle in his *Metaphysics* (Book I, Part 5, 350 BC),

> the so-called Pythagoreans [...] who [...] were the first to take up mathematics, not onlyari advanced this study, but also having been brought up in it they thought its principles were the principles of all things. Since of these principles numbers are by nature the first [...] all other things seemed in their whole nature to be modeled on numbers, and numbers seemed to be the first things in the whole of nature, [and] they supposed the elements of numbers to be the elements of all things, and the whole heaven to be a musical scale and a number.

The formulation of this position later given in Galileo's Assayer is well known (Il Saggiatore, 1632):

> Philosophy [i.e. physics] is written in this grand book – I mean the Universe – which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics […] without which it is humanly impossible to understand a single word of it; without these, one is wandering around in a dark labyrinth.

According to this view, objects have an intrinsic mathematical structure, independent of human perception or cognition; once a reference for comparison has been chosen, this position assumes then that measurement is a process aimed at discovering the values that properties of objects already and inherently have.[4]

The position that the purpose of measurement is to discover objectively-existing values of quantities extends well beyond the scientific community. For example, such a position helped motivate

[3] Under a Pythagorean conception that "numbers are in the world", another position would be that each property (or at least each quantitative property; see, e.g., Michell, 1999) of each object already is a value prior to measurement.

[4] If a property is specifically a quantity – which, again, is taken by this view (and consistently with the Aristotelian tradition) to be an empirical feature of the property itself, independently of the way in which it is modeled – the aforementioned reference for comparison would be a measurement unit, and the aforementioned process of discovery of values specializes as a discovery of *ratios of quantities* (see, e.g., Michell, 2004).

the introduction of the metric system during the French Revolution, as described by Witold Kula (1986: p. 123):

> The meter, in 'dehumanizing' measures, in rendering them independent of man and 'objective' in their interrelationship with man as well as morally neutral, has also transformed an instrument of 'man's inhumanity to man' into a means of understanding and cooperation for mankind.

Indeed, measurement results had been used in the past to give a social justification for decisions that otherwise could seem capricious; in the years following the first period of the French Revolution, many asked "what is the use to us of the abolition of the feudal system, if the *seigneurs* remain at liberty arbitrarily to increase or decrease their measures?" (p. 234). This refers to misuses of conventionality in the pre-Napoleonic choice of measurement units, at the same time implying that it must be possible to have just and fair measurement systems, and that a step towards this goal is to remove individuals' ability to arbitrarily "increase or decrease […] measures".

Yet even though the interpretation of measurement as aimed at the discovery of independently-existing values is attractive for its simplicity, it falls short precisely in that it equates measurement with a transmission process. Consider the nature of communication and transmission, as they are technically understood: communication is a mapping of one informational entity to another informational entity, performed through transmission, which is a mapping of one empirical entity to another empirical entity, e.g., electromagnetic waves (Shannon & Weaver, 1949). That is, the input to a communication system is an informational entity, explicitly provided by an agent who (or which) operates on purpose by encoding the informational entity into an empirical property of the transmission system, which is then transferred to the receiver and finally decoded into an informational entity. For example, a text may be encoded into an appropriately modulated amplitude of an electric signal, but also, trivially, into a sheet of paper with patterns of ink on it; the electric signal or the sheet of paper are transferred, and then decoded into a text. The informational entity transmitted along the transmission channel via the encoded empirical property is, at least in principle, perfectly knowable, given that it was purposely generated or selected by an agent. No such agent exists in the case of measurement, which in turn may be modeled as a mapping from an *empirical* entity to an informational entity: in a measurement process, the values of properties are the output, not the input. This essential difference between communication and measurement is depicted in Fig. 4.1.

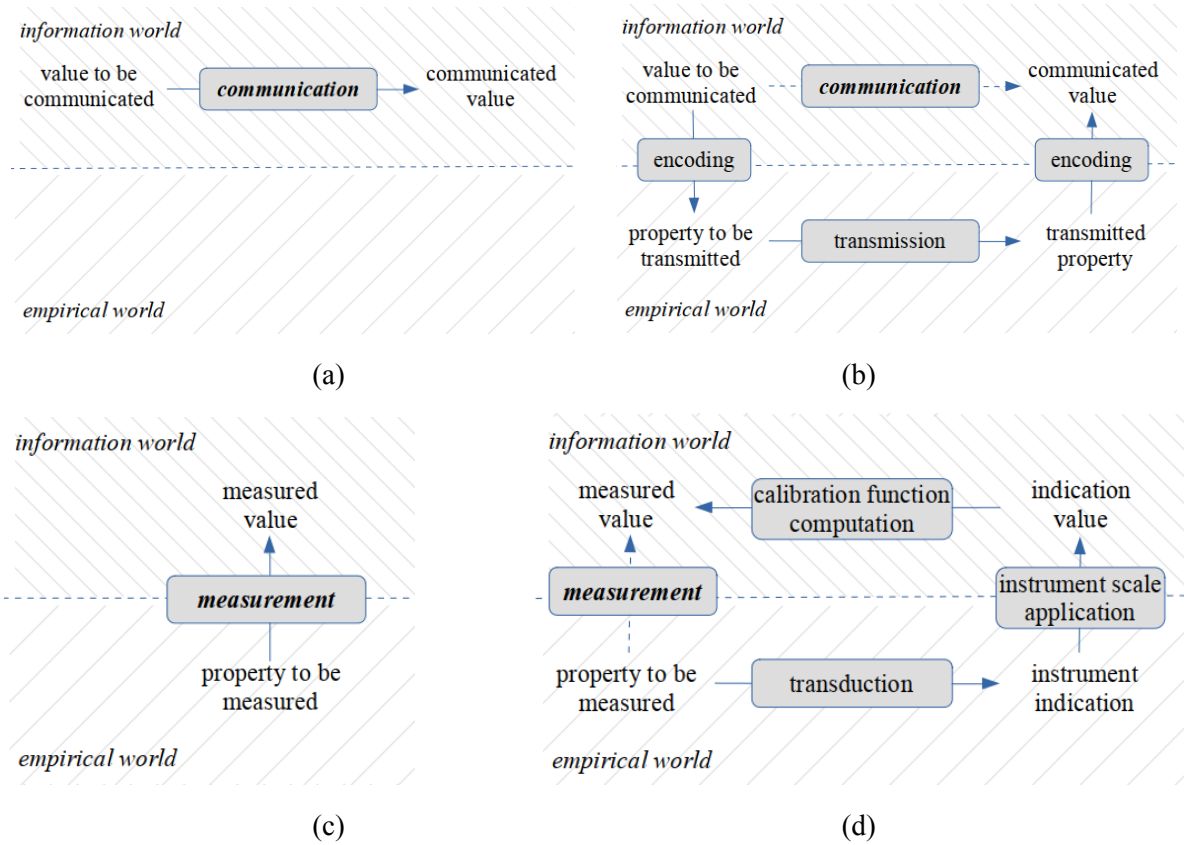


Fig. 4.1 A comparison between communication / transmission (black box (a) and open box (b) models), and measurement (black box (c) and open box (d) models, as elaborated from Fig. 2.10)

By neglecting this difference, naive realism misses some key features of measurement and is unable to account for the role of models in the definition of the measurand, in the development of the measurement procedure, and in the interpretation of the results of its application (e.g., Frigerio et al., 2010; McGrane & Maul, 2020; Teller, 2013).

Thus, a stereotyped (or naive) form of realism conflates the model with what is modeled and therefore measured properties and measured values.[5] Whenever the essential distinction between empirical and informational entities is absent, the simple position emerges that each property of each object has, inherently, a value, i.e., its "true value". The view that measurement aims at discovering such "true values" of properties can also be recognized in the conceptual foundation of the theory of measurement errors developed by Pierre Simon Laplace and Karl Friedrich Gauss at the beginning of the 19th century (see, e.g., Stigler, 1986; Rossi, 2014: ch.2), in which any quantity is assumed to have its own inherent (and in this sense "true") value, and the observed variability of the measurement results is explained as deriving from the introduction of errors. As discussed by Stephen Stigler (1986; 1992), the theory of errors was originally developed within the context of astronomy, wherein it seemed sensible to suppose that, for example, a planet really does have a (true) position, and therefore

[5] The usual notation "measured property of an object = measured value of a property" (which we have introduced as the *Basic Evaluation Equation* in Sect. 2.2.4) might contribute to such a confusion. As an example, consider the relation c = 299792458 m/s: does the symbol "c" stand for (a) the speed of light in vacuum or (b) its value? In the first case the relation conveys a claim about the physical world, which (were the metre defined independently of the speed of light in vacuum) is in principle true or false: the ratio of the speed of light in vacuum and the metre is 299792458. In the second case, the relation is just a conventional alias: "c" is synonymous with "299792458 m/s".

the idea that an observed position can be at least conceptually decomposed into a true position and error seemed straightforward:[6]

observation = truth + error

However, this reference to truth became murkier as the theory of errors was imported into the human sciences, in particular demography, psychology, and economics. As discussed by, for example, Stigler (1999) and Ian Hacking (1990), when in the 1830s the Belgian astronomer Adolphe Quetelet applied the theory of errors to measurements of the chest sizes of a sample of five thousand Scottish soldiers, he referred to the average (39.75") as an estimate of the "true" girth of a Scottish soldier's chest, a sort of Platonic ideal, and at times even appears to have ascribed to it a moral mandate. Shortly thereafter, Francis Galton (1869) found that the theory of errors could just as easily be applied to test scores as to chest sizes, now adding the idea that the average of a collection of observations (e.g., test scores, or individual item responses) about an individual person could be used to estimate something "true" of that person, an idea that was central to his studies of the inheritance of intelligence. But other scholars, like the philosopher and economist Francis Ysidro Edgeworth (1885), were less immediately convinced that such averages could be considered an estimate of something "true": as he put it,

> measurements by the reduction of which we ascertain a real time, number, [and] distance [are] cause[s], as [they] were the source from which diverging errors emanate. [...] Returns of prices, exports and imports, legitimate and illegitimate marriages or births and so forth, the averages of which constitute the premises of practical reasoning, are [...] descriptions. [...] In short, [the former] are different copies of one original; [the latter] are different originals affording one 'generic portrait'.

Indeed, as already discussed in Sect. 3.2.2, the concept of true value can be well defined mathematically; it is just a fact that, under well-defined conditions, sample means converge to the expected value of the underlying probability distribution. What is critical is the empirical interpretation of this convergence process and its limit point. A strongly realist perspective might suppose that the limit point reflects some kind of lawful feature of the world, and that, under the hypothesis of repeatability and the absence of biasing factors, the sampling process is required only to reveal it beyond experimental errors. But while such a perspective might have seemed relatively straightforward in the case of averaging estimates of the locations of planets, it is considerably less obvious what "feature of the world" is estimated by the average of a set of scores on test items, or in Edgeworth's other examples from the human sciences and social demography. As the theory of errors was increasingly applied outside astronomy, in contexts in which the measurement-independent existence of properties (and their values) is more controversial, explanations of it became increasingly divorced from metaphysically realist foundations; for example: "the true value [of the measurand] is the result that would be obtained by a perfect measurement. Since *perfect measurements are only imaginary, a true value is always indeterminate and unknown*" (Regtien, 2004: p. 44, emphasis added). In a similar vein, the development of Classical Test Theory (tellingly also referred to as "True Score Theory"; Lord & Novick, 1968) drew upon the mathematics of the theory of errors but rejected the realist metaphysics, instead formally defining the true score simply as the expected value of the raw score (which itself is, usually, the number of items answered correctly or endorsed by a

[6] Of course, the problem remains whether the number in the value of position is a real number with infinitely many significant digits, as a geometric model might imply. Were such a hypothesis to be maintained, and given that planets are not geometric points, the measurand should be changed to, e.g., the position of the center of mass of the planet. This would arise the new problem that, for the center of mass of a body to be uniquely defined, what is part of the body itself needs to be uniquely established, a condition that is hardly fulfilled by planets. Hence, the unavoidability of a non-null definitional uncertainty – as introduced in Sect. 3.2.4 – soon emerges also in these cases.

respondent out on a given test or survey) over a (hypothetical) infinite series of replications of administration of the test under identical conditions. As noted by Denny Borsboom (2005), while individual researchers might endorse realist interpretations of the true score (for example, as referring to the value of an existing quantity that is measured by the true score), its formal definition is more consistent with operationalism, as it is defined with reference to a particular test.

As conceptions of measurement became increasingly disconnected from assumptions about the metaphysics of properties, the way was opened to non-realist philosophical perspectives on measurement, as discussed further in the following sections.

The naive realist stereotype: measurement is analogous to a transmission process, which in the ideal case identically transfers the true value of the measurand to the measured value provided by the measuring instrument.

4.2.2 Operationalist perspectives on measurement

In the early part of the 20th century, as philosophical thinking was trending away from the naive forms of realism described in the previous section, philosophers who identified with the movement known as logical positivism synthesized many ideas from classical empiricism along with then-current advances in the philosophy of language and mathematics. Logical positivism was associated with the position that statements regarding unobservable (theoretical) entities should only be considered meaningful if they could be linked to observations in a clear and consistent manner.

The positivists saw measurement as a privileged means to establish the truth or falsehood of statements. From this perspective, the empirical sciences could delegate the responsibility of ascertaining the truth of their theories to measurement, as exemplified by the epistemic significance assigned to so-called "crucial experiments", which typically rely on high quality measurements. However, in contrast to the realist perspectives discussed previously, the positivists did not see numbers as being inherent properties of objects (Carnap, 1966: p. 100):

> Let us [...] consider the physical magnitude of weight. You pick up a stone. It is heavy. You compare it with another stone, a much lighter one. If you examine both stones, you will not come upon any numbers or find any discrete units that can be counted. The phenomenon itself contains nothing numerical – only your private sensations of weight. [...] We introduce the numerical concept of weight by setting up a procedure for measuring it. It is *we* who assign numbers to nature. The phenomena exhibit only qualities that we observe.

Since according to this perspective "the phenomena contain nothing numerical", the task of measurement is significantly recast (Carnap, 1966: p. 62):

> in order to give meaning to such terms as 'length' and 'temperature', we must have rules for the process of measuring. These rules are nothing other than rules that tell us how to assign a certain number to a certain body or process so we can say that this number represents the value of the magnitude for that body.

In the human sciences, behaviorism (e.g., Skinner, 1971) captured many of the same intuitions as those behind positivism, such as an emphasis on observables as the basis for science and an imperative to avoid metaphysical theories and concepts. In particular, the concept of the human mind was regarded

as too unobservable to be a proper object of scientific inquiry, as opposed to human behavior, which can be directly observed. Hence, due to its metaphysical grounds, the traditional form of realism discussed in the previous section was questioned, or simply rejected (Neurath et al., 1973):

> the statements of (critical) realism and idealism about the reality or non-reality of the external world and other minds are of a metaphysical character, because they are open to the same objections as are the statements of the old metaphysics: they are meaningless, because unverifiable and without content.

The positivists and behaviorists alike found a kindred spirit in the work of the physicist Percy Bridgman, who, influenced by many of the same cultural and historical factors that motivated positivism, had proposed a doctrine about the meaning of theoretical terms that came to be known as *operationalism* (or "operationism"): "we mean by any concept nothing more than a set of operations" (Bridgman, 1927: p. 5), with the corollary that as long as one has a "set of operations" (or "rules"), one has a measurement, simply by fiat. Thus, according to operationalism, the meaning of theoretical terms (including property terms) is exhausted by the particular operations undertaken to observe the entities designated by such terms. As Bridgman put it (p. 5),

> the concept of length is […] fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined.

Being a semantic (rather than methodological) doctrine, operationalism was not a theory of measurement *per se*, but led naturally to the characterization of measurement as simply "any precisely specified operation that yields a number" (Dingle, 1950: p. 11).

Operationalism had a strong influence on the human sciences (in particular, psychology) through the work of early behaviorists such as Burrhus Frederic Skinner (1945), and Edwin Garrigues Boring (e.g., 1923), who found it to be consistent with the prevailing empiricist attitudes of the time; the operationalist doctrine obviated the need to define measured properties independently of the manner in which they were assessed. Stanley Smith Stevens, who was Boring's student, famously asserted that "measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules" (Stevens, 1946: p. 677; see also see Leahey, 1980; Green, 1992; McGrane, 2015).

One obvious reason for the attractiveness of operationalism to early psychologists was the difficulty they faced in precisely defining psychosocial properties, which are not observable by traditional criteria, and seemed dangerously metaphysical in contrast to the positivist and behaviorist zeitgeist of the time: claims such as Boring's (1923: p. 35) that "intelligence is what the tests test" neatly sidestepped the issue. Notable in the idea of operationalism is that "it is meaningless to ask whether something is 'really' being measured [...] there is neither a need nor a place for postulating properties which are prior to the measurement operation" (Borsboom, 2005: p. 93). For example, according to an operationalist perspective, any form of knowledge or skill (e.g., reading comprehension ability) could be measured simply by assembling a set of questions that are judged (by whatever criteria) to be relevant; the property of reading comprehension ability could then be defined in terms of the answers a student gives to the set of questions. Stevens' definition of measurement is consistent with operationalism, insofar as the only given necessary condition for measurement is the presence of a rule (or operation) for numerical assignment. Stevens was indeed quite clear that this could be any rule other than "random assignment, for randomness amounts in effect to a nonrule" (Stevens, 1975: p. 47).

But despite the attractive simplicity of operationalist approaches to thinking about measurement, difficulties quickly become apparent. Measurement results, like many forms of information, are prized largely because of their *transferability*; by defining what is measured by a given instrument as just the property that interacted here-and-now with the measuring instrument, it is not obvious on what basis the information obtained could be used in reference to anything other than the current conditions.

Indeed, a coherent operationalist stance would hold that purely empirical means cannot demonstrate that different instruments measure the same property, given that any such demonstration unavoidably includes an inferential component. This has the consequence that each unique set of operations must be associated with a distinct concept; thus, for example, the outcome of the application of an alcohol thermometer and the outcome of the application of a mercury thermometer cannot refer to the same property, nor could two distinct tests of reading comprehension ability be claimed to measure the same property (Borsboom, 2005; Green, 2001). This is clearly inconsistent with common thinking and discourse; for example, most educators would immediately recognize that students' knowledge, skills, and abilities are not equivalent to their score on a particular test, and that such an identification narrows the definition of the property of interest to what usually would be considered just an *indicator* of the relevant properties.

For similar reasons, the very idea of measurement uncertainty is ill-fitting with operationalism: if a measured property is by definition what is measured by a given set of operations, what is there to be uncertain about? In particular, the idea that there may be uncertainty in the *definition* of the measurand – i.e., definitional uncertainty, the "component of measurement uncertainty resulting from the finite amount of detail in the definition of a measurand" (JCGM, 2012: 2.27; see Sect. 3.2.4) – is irreconcilable with operationalism, due to its inability to separate what is measured from how it is measured; one potential consequence of this is the misconception that what is true of the indication values is also automatically true of the measurand, "so that, for instance, properties are presumed to induce a linear ordering of people because [test scores] do" (Borsboom, 2006: p. 429).

For these reasons and others,[7] operationalism – at least in its original and radical formulation – has been almost uniformly rejected as irreconcilable with general scientific practice and vocabulary (Green, 2001; Bickhard, 2001), even by Bridgman himself ("I feel that I have created a Frankenstein, which has certainly got away from me", as he once lamented (Bridgman quoted in Frank, 1956: p. 75). The definition of <measurand> in the International Vocabulary of Measurement (VIM) – i.e., the "quantity *intended* to be measured" (JCGM, 2012: 2.3, emphasis added) – acknowledges the possibility of a distinction between the property with which the measuring instrument experimentally interacts and the property to which the produced values are attributed, and thus appropriately emphasizes that a model of the measurand is unavoidably present, though sometimes only implicitly, and that it is only through such a model that the produced information is transferable. Thus, operationalism, in and of itself, is simply incapable of explaining the societal role played by measurement, or of providing a justification for its acknowledged epistemic authority.

The operationalist stereotype: measurement is a model-free, purely procedural empirical process, which only provides information about the property with which the measuring instrument interacts.

[7] For a longer discussion of the history of operationalism, see, e.g., Chang (2019).

4.2.3 Representationalist perspectives on measurement

Many of the same socio-historical forces that helped shape the operationalist perspective – in particular, logical positivism, and more generally the strongly empiricist and antirealist attitudes prevalent throughout the early-to-mid 20th century – also helped shape what came to be known as the *representational theories of measurement* (RTM), which departed from and elaborated on Campbell's famous claim that "measurement is the process of assigning numbers to represent qualities; the object of measurement is to enable the powerful weapon of mathematical analysis to be applied to the subject matter of science" (1920: p. 267). The representationalists then formalized the idea of numerical (or, more generally, symbolic) assignment as the condition that defines measurement: for example, according to Patrick Suppes, an appropriate representation theorem "makes the theory of finite weak orderings a theory of measurement, because of its numerical representation" (2002: p. 59): even though weakly ordered entities do not in general satisfy the Euclidean conditions on a measure, they are considered measurable because they can be represented by means of numerical values.

The key concept here is *representability*, a condition formalized in terms of morphic mappings from empirical entities (either objects or their properties) to informational entities (Krantz et al., 1971; Narens & Luce, 1986).[8] According to this view, a precondition for measurement is the availability of some set of observed empirical relations amongst objects (e.g., x is greater than y and less than z), which are then mapped on to a symbolic system in such a way as to preserve qualities of their empirical relations. Consistently with positivist principles, this requires that empirical relations be directly observable, or "identifiable" (Suppes & Zinnes, 1963: p. 7), though it is not always obvious what this means (cf. Borsboom, 2005; Michell, 1990). Relational systems can possess different sorts of structures, and the particular sort of mapping of empirical to numerical relations determines the scale properties; for example, "x is greater than y" can be preserved by assigning to x any number higher than the number assigned to y, whereas "x is twice y" can be preserved by assigning to x a number twice that assigned to y. Such differences formed the basis of Stevens' system of scale types (most famously, nominal, ordinal, interval, and ratio; see Sect. 6.5), with the important consequence that, according to RTM, even non-quantitative properties could be considered measurable. Indeed, according to the representational perspective, "the question of measurement" is just "about the possibility of using numbers to describe certain phenomena", and by representational theories it "has received answers in the form of testable conditions" (Doignon, 1993: p. 473).

This purely formal interpretation of measurement is attractive for its epistemological simplicity, but the idea that measurement is definitionally equivalent to morphic mapping is so generic that it is unable to distinguish between measurement and consistent representation (Mari, 2013; Mari et al., 2012). Rather, what representational theories of measurement provide is an abstract framework for scale construction and meaningfulness of representation (Narens, 1985; 2002), and therefore at most for characterizing conditions of *measurability*. If, for example, a given quantity is selected as the unit, then the additive combination of that quantity with itself has to be associated with the numerical value 2, and so on. This is indeed a condition of morphic mapping which does not require any empirical action to be performed on given measurands, and is in fact preliminary to any such empirical action. In other words, scale construction is a critical precondition for the execution of measurement (see Sect. 7.3) but it is surely not measurement as such: representational theories take *as their starting point* the

[8] Whether such informational entities need to be numbers, with or without measurement units, is where Stevens, and since him representational theories of measurement, departed from Campbell. While, as just mentioned, numbers are required "to enable the powerful weapon of mathematical analysis to be applied to the subject matter of science" according to Campbell, Stevens (1946) made the representability of properties by means of numbers sufficient, but not necessary, for measurability.

availability of an observed set of relations among objects, and thus bracket out everything that must take place in order for such relations to be observed in the first place (see, e.g., Borsboom, 2005; Mari, 2013; Michell, 1990). Indeed, RTMs as such have little or nothing to say about the activities involved in data acquisition, including the definition of measured properties, the design and operation of measuring instruments, empirical and statistical strategies for controlling the effects of influence properties, and the management and reporting of measurement uncertainty. In the words of Marcel Boumans, himself referring to Michael Heidelberger, "The disadvantage of a general RTM is that it is much too liberal [...]: we could not make any difference between a theoretical determination of the value of a theoretical quantity and the actual measurement." (2007: p. 234). Instead, RTMs might be better interpreted as a purely formal and idealized interpretation of measurement (sometimes even explicitly noted in accounts of representational theories, e.g., "the theory of measurement is difficult enough without bringing in the theory of making measurements", Kyburg, 1984: p. 7), but as such are unable to distinguish between measurement and morphic mapping in general (Mari, 2013; Mari et al., 2012). From this perspective one could argue that the representational theories of measurement are simply misnamed: they are at most a theory of scale construction, presupposing the availability of the right sorts of empirical inputs to form the basis of the resulting scale. A better term for them might be then *representational theories of scaling*.

Representationalism also has the consequence that evaluations based on orderings and even classifications count as measurements (related to what Stevens called "ordinal" and "nominal" scales respectively), and therefore introduced a multiplicity of algebraic structures in which the properties and their values can be embedded. Specific to each scale type is the set of relations that are invariant under particular scale transformations.[9] The acknowledgment that such structures are not inherent features of properties, but instead depend on the state of knowledge about that property (Giordani & Mari, 2012)[10], can be interpreted as an attempt to give measurement an epistemic foundation, rather than an ontic foundation such as was proposed by the realists discussed previously.

Finally, the representationalist emphasis on rules also had significant consequences for the concept of a true value. While a rule (e.g., for numerical assignment) could be found to be (for example) adequate, effective, or efficient, its application is in and of itself unrelated to the possible *truth* of its outcomes. But by the time of the introduction of RTMs, the idea that measurement is a quest for true values had become so entrenched in measurement science that renouncing it appeared to be unacceptable, even in a context in which the search for an ontic grounding for the concept of a true value had been replaced by epistemic (or even formal) conditions. The definition of <true value> given in the VIM exemplifies this. According to the first edition (ISO, 1984: 1.18) a true value is "the value which characterizes a quantity perfectly defined, in the conditions which exist when that quantity is considered" (though what would count as a perfect definition of <quantity> is not explicated). Almost thirty years later, the third edition of the VIM changed the definition: a true value is a "quantity value consistent with the definition of a quantity" (JCGM, 2012: 2.11). The consistency of something with something else is a condition that can be obtained by the appropriate application of a rule, but consistency and truth are distinct concepts, and this seems to be a definition of <consistent

[9] Stevens' theory is not without objections (for one synthesis of criticisms, see Velleman & Wilkinson, 1993). In part, such criticisms have reacted to Stevens' choice of calling the invariant scale transformations "admissible" or "permissible", the objection being, in essence, that research should not be driven by prescriptions and surely not inhibited by proscriptions. Our analysis of these criticisms is in Sect. 6.5.1.

[10] For example, temperature, thought in antiquity to be an ordinal property, was upgraded (with the introduction of thermometers and thermometric scales) to a quantity. At first only interval-level measurement was possible; eventually the thermodynamic re-definition of temperature, which introduced a non-conventional zero point in the scale, made ratio-level measurement possible.

value>, rather than of <true value>. Thus truth has been maintained in the lexicon of the VIM but seems to have disappeared in the substance.

Representationalism has had relatively little direct impact on the practices of either the physical or the human sciences. This fact can be interpreted as a sign of the practical uselessness of such theories in situations – like most cases of physical measurement, and many cases of non-physical measurement as well (see, e.g., Cliff, 1992) – in which the source of complexity, and hence of interest, is actually (also) the execution of measurement, not (only) the characterization of its preconditions. However, as already noted, representationalism has had at least an indirect impact on thinking about measurement through the work of Stevens, who, informed by both operationalism and early versions of representationalism, defined measurement "the assignment of numerals to objects according to a rule" (Stevens, 1946: p. 667). This definition and close variants thereof are ubiquitous in introductory textbooks on psychology and psychological statistics (for a review, see, e.g., Michell, 1997), when indeed a definition of measurement can be found at all (see also Borsboom, 2009).

Stevens proposed this definition after the Ferguson Committee, which was initially convened in 1930 by the British Association for the Advancement of Science (Ferguson et al., 1940) with the charge of studying the possibility of providing "quantitative estimates of sensory events", ultimately concluded that claims of measurement being made by human scientists of the day – including by Stevens himself, in the context of his work on the measurement of sensations – were at best premature, and at worst "not merely false, but misleading" (p. 345) given the way in which measurement was understood by the wider scientific community (for longer histories than is possible here, see, e.g., McGrane, 2015; Michell, 1999; Rossi, 2007; see also Sect. 6.5). Stevens' move of redefining measurement arguably permitted psychological scientists to avoid addressing the criticisms of the committee, and gave them license to continue making claims of measurement – along with at least implicit claims to the epistemic authority and social status that measurement had earned amongst the general public – and had ripple effects that effectively unmoored the development of the theory and practice of measurement in the psychological sciences from the rest of science and philosophy (McGrane, 2015).

But Stevens' definition of measurement inherited the limitations of both operationalism and representationalism, and in effect trivialized the requirements for claiming that one has successfully achieved measurement; to abbreviate an example from Borsboom (2009), one could divide individuals' shoe sizes by their postal codes and assign to them the resulting numerals; according to Stevens' definition, this would be an instance of measurement, though it would be difficult to state what, if anything, is being measured. Viewed this way, the identification of a given process as an instance of measurement is simply unrelated to the issue of whether its results are authoritative or worthy of trust. Thus representationalism, like operationalism, is incapable of explaining the societal role played by measurement, or of providing a justification for its acknowledged epistemic authority.

The representationalist stereotype: measurement is any process that maps properties to informational entities in a consistent way, so as to represent the former by means of the latter.

4.3 The concept of validity in psychosocial measurement

As has been described in the previous sections, the way in which measurement is commonly defined in the human sciences – i.e., as rule-based numerical assignment – is, in and of itself, not

usable as a basis for justifying the epistemic authority of measurement processes and their results. Additionally, as was discussed in Chap. 1, the human sciences generally lack anything like the systems of lawful relationships (sometimes referred to as "nomic nets", or by the VIM as "systems of quantities"; JCGM, 2012: 1.3) that serve as the basis for identifying transduction effects that themselves serve as the basis for physical measurement processes (see also Finkelstein, 2003; 2005), and thus also serve as justifications for claims regarding the existence of measured properties (see also Sect. 6.6) and the validity of measurement processes involving these properties and the relevant laws. However, this does not mean that human scientists have been unconcerned with such justifications; rather, particularly in the contexts of educational testing and psychological research, a separate concept (and literature) has arisen that addresses the issues of when and why some measurement results are (or are thought to be) epistemically authoritative in their role of providing information about a property intended to be measured: that is, the concept of *validity*, which is often described as "the most fundamental consideration in developing and evaluating tests" in the human sciences (quote from the *Standards for Educational and Psychological Testing*, AERA, 2014: p. 11; for recent comprehensive treatments of the topic of validity from philosophical perspectives, see Markus & Borsboom, 2013, and Slaney, 2016). Moreover, as has been argued by, e.g., Borsboom (2006; 2009), Michell (2009), Maul et al. (2016), and Slaney (2017), the actual practice of psychosocial measurement seems to be largely disconnected from and unconcerned with the philosophical conceptions of measurement described in this chapter; indeed, within the mainstream literature on psychosocial measurement, one would be hard-pressed to find serious engagement with even basic philosophical questions such as what measurement is, even in authoritative sources such as the previously-mentioned *Standards* (Maul, 2014; see also Borsboom, 2009; Michell, 1997). Conversely, the literature on validity has developed more directly in tandem with the practice of psychosocial measurement (Newton & Shaw, 2014), and thus is both more reactive to and influential on such practice. Thus, to more thoroughly appreciate how thinking about measurement has developed in the human sciences – also in the service of our larger goal of understanding measurement across the sciences – it will be useful to briefly review the literature on validity and validation.

There have been several distinct phases in the history of thinking and discourse about validity over the 20th and 21st centuries, often dovetailing with the history of thinking and discourse about measurement (as reviewed in the previous sections), and science and knowledge even more generally. Even today, despite wide agreement regarding its importance, there is no single conception of validity universally accepted in the scholarly and professional communities in the human sciences, and there remains considerable controversy surrounding its definition, as well as about many related concepts and terms. This next section briefly reviews this history (see also Maul, 2018).

4.3.1 Early perspectives on validity

One of the earliest proposed conceptions of validity, and one that remains popular among many contemporary scholars and practitioners, is that validity is about the extent to which an instrument measures what it is claimed to measure.[11] Against the backdrop of logical positivism, behaviorism, and operationalism (as discussed in the previous sections), formal accounts of validity in the early 20th century – due to scholars such as Truman L. Kelley and Edward E. Cureton – operationalized the idea

[11] Oftentimes validity is introduced alongside the concept of *reliability*, which usually refers to the extent to which measurement results are free from random sources of measurement error. Although some sources (e.g., Moss, 1994) describe reliability and validity as separate, complementary issues, most contemporary descriptions emphasize that reliability is a precondition for validity rather than a separate issue. As was discussed in Sect. 3.2.1, this usage of the terms "reliability" and "validity" then seems to map fairly closely onto what the VIM refers to as "precision" and "accuracy" respectively (JCGM, 2012: 2.15 and 2.13).

of an instrument measuring what it is claimed to measure in terms of the statistical correlation between test scores (and therefore more generally instrument indications, to use the metrological terminology) and a "criterion variable", that is, an external criterion that was believed to be related in some way to the property that the test is measuring. For example, the validity of a job placement test might be defined in terms of the statistical correlation between its scores and some kind of quantified information about job performance, or a short version of a test might be evaluated in terms of the correlation of its scores with those of a longer or more thorough battery of tests. Such test-criterion correlation coefficients were sometimes referred to as *validity coefficients*.[12]

In other contexts, tests are developed from sets of content specifications: for example, for many educational tests, a primary goal is to ensure adequate instructional attention to a domain covered in a course. In such contexts, prediction of a specific external criterion could be regarded as less important than ensuring that the content of the test was representatively sampled from the domain of interest; this, in turn, is primarily established via documentation of the test construction procedures and through expert review. This led to a distinction between *criterion-related validity* (also sometimes called *predictive validity*) and *content validity*, initially thought of as each applying to different types of tests.

4.3.2 Construct validity

Starting in the 1950s, scholars such as Lee Cronbach and Paul Meehl began to observe that while the concepts of criterion validity and content validity seemed to be appropriate for many tests, some other kinds of tests appeared to require something else. In particular, psychological properties such as personality characteristics (e.g., aggression, conscientiousness) and broadly defined cognitive abilities (e.g., general intelligence) seemed difficult to operationalize in terms either of a specific domain of content coverage or in terms of relations with specific external criteria. This led Cronbach and Meehl (1955) to introduce the concept of *construct validity*, which was understood primarily in terms of how the property measured by a given test related to a network of other properties, that is, a *nomic network* (which they called a *nomological network*), which could in turn be estimated by examining, under specified conditions, the statistical correlations between scores on the test and scores on other tests (or other quantified information about the relevant properties); one could then examine the extent to which these correlations were consistent with predictions made based on the theory of what the test measured.[13]

One consequence of the focus on construct validity was increased attention to the distinction between a test and the psychological property the test was designed to measure, oftentimes referred to as a *construct*[14] (as opposed to the earlier, operationalist view that a test simply defined a property).

[12] The influence of operationalism seems clear: while today one might still speak of a correlation coefficient as a *tool* for the evaluation of validity, speaking of such a correlation as *definitional* of validity blurs the distinction between what we know and how we know it. See also Borsboom et al. (2004).

[13] Cronbach and Meehl's conception of nomological networks drew from Carnap's (1950) project to specify how theoretical terms are defined implicitly through the role they play in networks of lawful relations. However, as discussed briefly in Chap. 1, this project did not work for the simple reason that "there were (and are) no nomological networks involving concepts like general intelligence" (Borsboom et al., 2009: p. 136).

[14] Although many sources treat (or appear to treat) the term "construct" as synonymous with "property" (or at least "psychosocial property"), other sources also use it to refer to a concept or linguistic label that *refers* to a property (for a discussion, see, e.g., Slaney & Racine, 2013), and some sources even do both simultaneously: for example, the *Standards for Educational and Psychological Testing* (AERA, 2014), which are discussed further below, define a construct as "the concept or characteristic that a test is designed to measure" (p. 11). To avoid confusion, we use the terms "property" and "concept of property" rather than "construct", except when specifically referring to language used by others. We also discuss in Sect. 4.5 properties that are in some sense *constructed* by human minds or human activities. Perhaps jarringly, in the terminology common in the human sciences (and construct validity theory in particular), the term "construct" might or might not imply that the property is thought of as having been constructed (see, e.g., Slaney,

This raised awareness of the possibility that multiple tests could measure the same property. Donald T. Campbell and Donald W. Fiske (1959) popularized the idea of multimethod studies, and in so doing added two new terms into the validity lexicon: *convergent validity*, reflecting the idea that the results of measurements made with different instruments but of the same property should exhibit high levels of agreement with one another (i.e., they should "converge" on a common outcome), and *discriminant validity*, reflecting the idea that the results of measurements of different properties should not be too highly correlated with one another, even if they used similar kinds of instrumentation (i.e., it should be possible to empirically "discriminate" among theoretically distinct properties). As a classic example, if extroversion and dominance are both assessed via self-report and the reports of one's family members, evidence for convergent validity could take the form of showing that self-reports and family reports of extroversion are highly correlated (and similarly for dominance), and evidence for discriminant validity could take the form of showing that self-reports of extroversion and self-reports of dominance are not so highly associated as to render them empirically redundant (and similarly for family reports).

4.3.3 An argument-based approach to validity

Starting in 1989, Samuel Messick offered a new perspective on validity that reflected a significant shift from previous viewpoints in two important respects. First, Messick's view subsumed disparate lines of validity-related evidence under a generalized version of construct validity; thus, according to this view, (construct) validity is a single property of a test. Second, Messick reframed validity from being a claim about the true state of affairs ("a test measures what it claims to measure") to being a claim about the present state of available evidence, as judged by a particular community – that is, from an *ontological* claim to an *epistemic* claim: "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989: p. 13). The idea of distinct types of validity (e.g., criterion, content, construct) was replaced with the notion of there being distinct types of *evidence* that could be brought to bear on the validity of a given test, depending on the intended purposes of the test. Broadly, these types of evidence help establish that the test assesses as much as possible of what it should assess and as little as possible of what it should not: in Messick's language, this involves minimizing both *construct underrepresentation* and *construct-irrelevant variance*.[15]

One of the more controversial elements of Messick's perspective was the proposition that validation should explicitly involve a consideration of the consequences of test interpretation and use. For example, if educational tests given to students are used to help inform decisions about the retention and compensation of teachers, claiming that the tests are valid for this purpose would involve demonstrating not only that they measure the knowledge, skills, and abilities of students they claim to measure, but also that using these tests as a basis for high-stakes decisions about teachers has the intended positive consequences and does not have unforeseen negative consequences. This viewpoint could be taken as broadening the concept of validity to include social and moral concerns in addition to more purely epistemic concerns. Although Messick himself only proposed that the consequences of tests could be used as indirect evidence of construct underrepresentation and construct-irrelevant variance, other scholars such as Lorrie Shepard (e.g., 1993) made stronger proposals for the explicit consideration of consequences as a primary and independent source of validity evidence.

Messick's view of validity has remained influential since its introduction, and is arguably still the dominant conception of validity in the literature on educational assessment and measurement. Using

2016).

[15] The term "influence properties" could be thought of as referring to sources of construct- (or property-) irrelevant variance.

Messick's definition of validity as a starting point, scholars such as Michael Kane (e.g., 1992) have argued that the activity of validation should consist of the construction and evaluation of an argument (or set of arguments) aimed at defending the appropriateness of a test for a particular, well-specified use; the specification of such an argument then serves as an organizing framework for the collection of forms of evidence necessary for its defense. Kane's argument-based approach is targeted to the practical problem of validation and not a new theory about validity itself; this emphasis on validation rather than validity reflects a shift in focus toward pragmatic, context-specific arguments tailored for specific audiences and circumstances. The argument-based approach emphasizes that any validation effort begins with a clear statement of the proposed uses and interpretations of a test, and that if tests are used for purposes other than those originally intended, this requires a reexamination of the argument or the development of an entirely new argument. According to this view, the consequences of testing would play a central role in a validity argument for a given test insofar as the proposed use of the test implies an intention for certain consequences to happen (or not to happen) as a result.

Messick's and Kane's perspectives have been influential in shaping recent editions of the *Standards for Educational and Psychological Testing* (e.g., AERA, 2014; see also Wilson, 2005) which aim to provide guidance to practitioners on the construction of convincing arguments for the adequacy and appropriateness of tests for given purposes, and describe different types of evidence that might be brought to bear on such arguments. Echoing Messick, the *Standards* define <validity> as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests". The *Standards* go on to specify that sources of such evidence may come from several sources, related in particular to (a) the content of the test (evaluated by, e.g., expert judgment of the alignment of test items and the property, the content and formats of the items, and the materials supporting test interpretations); (b) the cognitive processes engaged in by examinees when responding to test items (evaluated by, e.g., interviews, observation, or self-report), (c) the extent to which empirical patterns of test results are consistent with theory-based expectations; (d) the extent to which patterns of empirical relations between the test results and other forms of quantified information are consistent with theory-based expectations; and (e) patterns of real-world outcomes ("consequences").

Of course, the degree of support for any proposition is logically independent of the truth of that proposition, and thus the conception of validity could be regarded as essentially *legalistic* rather than scientific.[16] Although the term "measurement" is used throughout the Standards, it is never defined; it seems to be used interchangeably with the terms "testing" and "assessment". Thus, as the concept of validity has expanded, the concept of measurement has arguably been buried, and it is not always obvious how or even if they are intended to relate (for an expanded discussion, see Maul, 2014).

It could be noted that interpreting tests or assessments as measuring instruments is only one among many possible interpretations, and tests are routinely put to uses that, strictly speaking, do not appear to require that any measurement take place at all; for example, the function of a school exam might simply be to inspire students to study (which is sometimes termed a *signification* purpose of a test; Wilson, 2018). Thus, it could be argued that the interpretation of validity as being about whether a test measures what it claims to measure is a special case of a broader focus on test score interpretations and uses, as advocated by Messick and Kane. But especially insofar as many test interpretations and uses do at least appear to depend on measurement claims, it is still necessary to articulate and justify claims about measurement separately and in addition to other claims about test interpretation and use more broadly.

[16] A legalistic conception of validity "operationalizes the concept in a way that makes it clear for test developers what the exact standard for validity is: they have to convince the jury. This bears all the marks of a licensing procedure. However, for scientific research, licensing procedures do not suffice. Truth cannot be […] equated to amounts of evidence" as noted by Borsboom (2012: p. 40).

4.3.4 Causal perspectives on validity

In contrast to the perspective of Messick, Kane, and the *Standards*, other recent scholarship on validity has more strongly emphasized understanding its semantics in terms of (a) factual claims about true states of affairs, rather than judgments based on available evidence, and (b) measurement, rather than interpretations and uses more broadly. In particular, Borsboom and colleagues have developed an account of validity that could be regarded as an extension of the earliest definition of the term (i.e., validity is whether a test measures what it claims to measure): specifically, "a test is a valid [measuring instrument] of a [property] if (a) the [property] exists and (b) variation in the [property] causes variation in the outcomes of the test" (Borsboom et al., 2004). This perspective on validity emphasizes that whether or not a test is valid as a measuring instrument of a property is a claim about the state of affairs in the world, and its truth or falsity is independent of the evidence available at any given time, or the extent to which that evidence is found to be persuasive by any given community of observers.[17]

Borsboom and colleagues' perspective on validity is also the most compatible with the framework presented in this volume, where validity can be understood (using terminology to be defined more precisely in later chapters) in terms of the distinction between the *intended* and *effective property* measured by an instrument, with ideal validity being definable as a perfect union between the two.

However, it seems fair to say that Borsboom and colleagues' view still stands outside the mainstream of thinking and discourse about validity (see, e.g., Newton, 2012, for a discussion). The dominant trends in thinking about validity over the 20th and early 21st centuries appear to roughly follow at least some aspects of the historical progression of thinking about measurement found in the philosophical and metrological literatures (and as described in previous sections of this chapter), moving from an early form of realism ("whether a test measures what it claims to measure") to antirealist or nonrealist stances influenced by logical positivism (nomological networks) and pragmatism (adequacy and appropriateness for an intended purpose), and now possibly back to a form of realism, albeit one that acknowledges the role of models both of and in the measurement process. This last shift helps point the way towards the philosophical perspective that grounds the theories and models of measurement developed in the balance of this volume, which we discuss further in Sect. 4.5. Prior to this, however, we will explore some of the key consequences of the various philosophical perspectives we have discussed so far as a way of highlighting what is at stake with respect to characterizing measurement in terms of its sufficient conditions.

4.4 An interpretive framework

As described in the previous sections, the stereotyped versions of naive realism, operationalism, and representationalism share a common presupposition: measurement can be understood as a black box that somehow transforms inputs into outputs. In Chap. 2 we adopted a top-down strategy for progressively characterizing measurement in terms of more and more specific *necessary* conditions, which allowed us to leave the box closed. But, as previously discussed (particularly in Sect. 2.2.1) measurement is an empirical process, and complementary *sufficient* conditions should be able to explain the epistemic role customarily attributed to measurement and its results. It seems plausible, then, that these sufficient conditions should be characterized *in terms of the structure of the process*, thus by "opening the box", and with the desirable meta-condition that such a structural

[17] Consistently with this perspective, Wilson (2005) has advocated that instrument development efforts in the human sciences focus on the development of the definition of the property, and then the specification of theory regarding how this property is related to test outcomes. This perspective is explored further in Wilson (2013), and Chap. 7.

characterization should apply to the measurement of both physical and psychosocial properties. While we leave a thorough exploration of what is "inside the box" to Chap. 7, it can already be acknowledged that what we will seek "inside the box" is informed by an underlying philosophical standpoint, and therefore that the distinct positions on measurement described so far leave open that there can be more than one way to answer the question of what measurement is even in the case of black box models. With the purpose of offering a structured perspective about this multiplicity, we introduce here a simple framework driven by two basic questions (Mari, 2013).

Q1. Are *empirical* constraints on the process relevant for the definition of <measurement>; i.e., should the definition include reference to any empirical conditions?

An affirmative response to this question means that only under some specific conditions regarding the way the measurement process is empirically performed[18] is an evaluation to be considered a measurement; a negative response means that empirical constraints are immaterial for characterizing what is measurement.

Q2. Are *informational,* and more specifically *mathematical* constraints on the process or the measured entities relevant for the definition of <measurement>; i.e., should the definition include reference to any mathematical conditions?

An affirmative response to this question means that only if the measurement process fulfills some mathematical conditions or is applied to entities fulfilling some mathematical conditions[19] is an evaluation to be considered a measurement; a negative response means that mathematical constraints are immaterial for characterizing measurement.

Given our preliminary condition that measurement is a property evaluation (see the related discussion in Chap. 2),

- Q1 prompts an investigation into the conditions sufficient to identify measurement *directly through the structure of the process*: if it is not the case that every evaluation is a measurement, how is measurement specified? And
- Q2 prompts an investigation into the conditions sufficient to identify measurement *indirectly through the structure of (i) measurable properties or (ii) measured values*: if it is not the case that every property is measurable, how can we specify which properties are measurable?

These questions do not in principle relate to the physical or non-physical nature of the property to be measured. Of course, further dimensions might be added to make the framework more specific, but any standpoint on measurement has to account for its position with respect to Q1 and Q2, which in principle may be treated as distinct and independent criteria. For the sake of simplicity both Q1 and Q2 are phrased as Boolean questions, thus bracketing out the possibility of intermediate positions: empirical conditions (Q1) are considered either relevant or not, and mathematical conditions (Q2) are considered either strictly relevant or not. Accordingly, four general positions are then identified, depending on whether one considers the conditions relevant for the definition of <measurement> to be:

α: mathematical but not empirical, or

β: both empirical and mathematical, or

γ: neither empirical nor mathematical, or

δ: empirical but not mathematical.

[18] See for example the answer that in Table 4.1 is given to the question "is measurement characterized by the structure of the process?".

[19] For example, a traditional condition might be the invariance of ratios of properties.

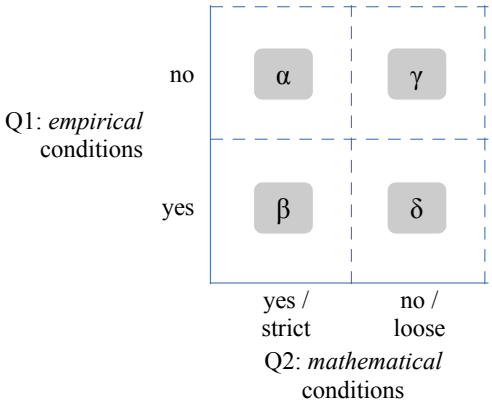


Fig. 4.2 A simple framework for mapping conceptual perspectives on measurement

As depicted in Fig. 4.2, this option space is a partially ordered set, where γ is the least constraining position and β is the most demanding one. The perspectives on measurement considered in previous sections of this chapter may be (partially) understood by examining how they would address Q1 and Q2: let us explore this option space.

4.4.1 Exploring perspectives on measurement

Euclid's Elements (Euclid, 2008) set the stage by taking geometry to be the paradigm of measurement: "a magnitude is a part of a(nother) magnitude, the less of the greater, when it measures the greater; the greater is a multiple of the less when it is measured by the less; a ratio is a sort of relation in respect of size between two magnitudes of the same kind" (Book 5, definitions 1–3). However, as already discussed in Sect. 3.4.2, what is at stake with this concept of <measure> (note, not measure*ment*) was clearly pointed out by Augustus De Morgan: "the term 'measure' is used conversely to 'multiple'; [...] hence [if] A and B have a common measure [they] are said to be commensurable" (1836: p. 9).[20] Not surprisingly, then, in this context the English term "measurement" (also written "mensuration" in the past: see, e.g., Hutton, 1795) mainly refers to procedural demonstrations of geometric propositions, such as "the area of any circle is equal to a right-angled triangle in which one of the sides about the right angle is equal to the radius, and the other to the circumference, of the circle", as taken from Archimedes' short treatise titled "Measurement of a circle" (Heath, 1897: p. 91-98). Of course, as discussed in the previous chapter, no empirical activities are expected here, or even allowed, according to Euclid: "in the geometrical constructions employed in the Elements [...] empirical proofs by means of measurement are strictly forbidden" (Fitzpatrick, 2008; in his introductory notes to his translation of Euclid's Elements). Hence, this is the original case of position α (see Fig. 4.3), which may be summarized as: *measurement is quantification* (see also Sect. 3.4.2, where it is argued that this position may be understood as based on the assumption that <measurement> and <measure> are identified). Today, this position might be seen as properly related to a branch of mathematics, as opposed to empirical science, but the historical labels tend to confuse this conceptual separation.

[20] This is indeed the Euclidean position: *x* measures *y* if *y* is a multiple of *x*.

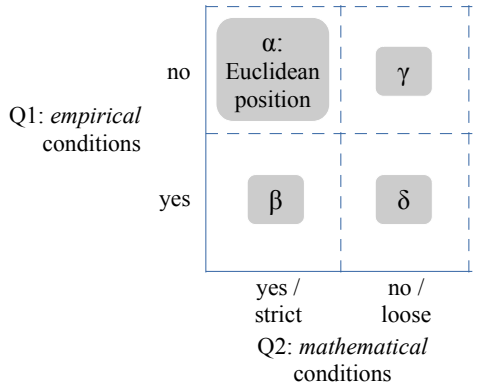


Fig. 4.3 The starting point: the Euclidean position in the framework

Centuries later, in the context of the adoption of the experimental method, the Galilean motto of "measuring what is measurable and making measurable what is not yet" was meant primarily as a call for application of empirical methods and innovation in instrumentation, an attitude that has been interpreted as sharply discontinuous with earlier traditions. For example, in reference to the functioning of science before Galileo, Alexandre Koyré stated that "no one had the idea of counting, of weighing and of measuring; or, more exactly, no one ever sought to get beyond the practical uses of number, weight, measure in the imprecision of everyday life" (1948). The Euclidean characterization of measure was maintained, but complemented with an interest in discovering physical transduction effects (see Sect. 2.3) and designing and producing devices that implement them. Such an emphasis on empirical activities was very effective in "making measurable" properties that had never been measured before, such as pressure and temperature, and later electrical and magnetic quantities. This triggered a new interest for scientists, and more specifically physicists (who were called "natural philosophers" at the time of Galileo and Newton), to develop instrumentation, on their own or through novel collaborations with craftsmen and then engineers. This corresponds to position β in Fig. 4.4.

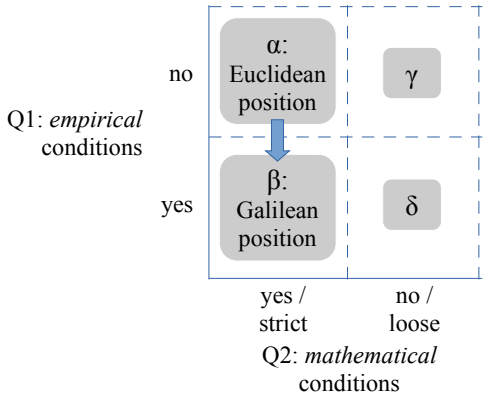


Fig. 4.4 The first transition: the Galilean position in the framework

As mentioned above, β is the most demanding position, as it inherits mathematical constraints from Euclid (i.e., a ratio must be defined between properties for them to be measurable), and empirical constraints from Galileo (i.e., properties must be observable, either directly or through their transduction to observable properties, for them to be measurable). But while these constraints may have seemed reasonable for the measurement of properties that Campbell would later refer to as *extensive* – i.e., properties that could be empirically concatenated, such as spatial distance, mass, and volume, and which could therefore be demonstrated to physically satisfy the Euclidean requirement of

property-related additive divisibility – they became a matter of controversy with respect to the so-called *intensive quantities*, like temperature and density, and even more so with non-physical properties such as those of interest to psychophysicists (i.e., intensity of sensations) like Gustav Fechner in the late 19th century (see, e.g., Gescheider, 2013).[21]

Difficulties with issues such as these were a major part of the motivation for the formation of the Ferguson committee, as described previously, which was charged with studying the possibility of providing "quantitative estimates of sensory events". The committee basically endorsed the Euclidean perspective on additive divisibility as a necessary condition of measurability, with the consequence that "the main point against [for example] the measurability of the intensity of a sensation was the impossibility of satisfactorily defining an addition operation for it" (Rossi, 2007: p. 551; see also Sect. 6.5). Indeed, additivity is at the basis of what Campbell called "fundamental measurement" (1920: p. 267), from which, in his account, any other form of measurement needs to be derived, corresponding to the possibility of obtaining an intensive quantity (e.g., density) as a function of extensive quantities (e.g., mass and volume).

Of course, in general, non-physical properties are not empirically additive. In the face of the possible conclusion that this simply precludes the possibility that such properties could ever be measured, two paths were explored.

The first path started from the observation that if a non-additive quantity like density is acknowledged to be measurable, it is because even in the Euclidean context empirical additivity is not necessary for measurability. Rather, what is required is the possibility for one to meaningfully interpret the relation $Q[a] = x\ q_{ref}$ (see Sect. 2.2.4), for example $L[a] = 1.2345$ m, as a ratio of the two involved properties – the length of the object a and the metre – i.e., $x = Q[a] / q_{ref}$.[22] Hence, the critical condition is the provision of a meaningful interpretation of the ratio of individual properties of the same kind (Rossi & Crenna, 2013): this can be granted by the empirical additivity of some quantities, and may be derived for quantities which are functions of additive quantities, but in principle could also be obtained in other ways. This is the strategy, in particular, that led to the development of so-called *additive conjoint measurement* (Luce & Tukey, 1964), and is arguably also at the core of the Rasch approach to measurement (Rasch, 1960; see, e.g., Borsboom, 2005: ch.4; Wilson, 2013). In our structural perspective this path remains in position β, and therefore could be characterized as a constructive approach for embedding non-physical properties into the Galilean conception of measurement.

The other path was triggered by the emphasis on the representational role of measurement, thus with an emphasis on numerical *assignment* rather than *determination* (Mari, 1997). By conceiving of numbers in the Euclidean sense of ratios of quantities, Campbell's claim that "measurement is the process of *assigning* numbers to represent qualities" (1920: p. 267, emphasis added) was still conservatively bound to the algebraic condition that only properties that admit of ratio- or interval-level representation are measurable. But this was also the starting point for another interpretation,

[21] As in definition 1 of Book 5 of the Elements, as previously quoted, the condition for a quantity (a "magnitude" in the traditional translation) "to measure" another quantity is that the first *is a part of* the second: "a magnitude is a part of a(nother) magnitude, the less of the greater, when it measures the greater" (Euclid, 2008). But for a property P which makes objects a, b, ..., comparable through an order relation (or least as a partial order), it is clear that $P[a]<P[b]$ does not generally mean that the property P of a is a part of the property of b. Indeed, it is additivity that guarantees this condition.

[22] This condition can be generalized by admitting that sometimes the zero of Q is not an intrinsic feature of Q, so that the numerical value x in the relation $Q[a] = x\ q_{ref}$ is determined only when a zero property q_0 is set for Q, as $x = (Q[a] - q_0) / (q_{ref} - q_0)$ (for example, this was the case of temperature before the introduction of thermodynamic temperature and its measurement in kelvins, and is the case of position along a line, which, differently from length, can be measured only having chosen a reference/zero position). Since, in most cases, non-physical properties do not have an intrinsic or obvious zero, this generalization – leading to what Stevens called an "interval scale" (1946) – proved to be very important for the development of measurability conditions for non-physical properties.

according to which the important point is not representation by means of numbers, but representation as such. As discussed in Sect. 4.2.3 this standpoint was developed in particular by Stevens, who accepted measurement as representation by means of informational entities (which he called "numerals"), instead of numbers only, and introduced a condition of consistency in the assignment that he called "permissibility", closely related to what was then referred to as "meaningfulness" (Narens, 2002): the relations observed among measured properties must also apply among the assigned values, and – most importantly – only the informational relations corresponding to empirical relations should be exploited in inference and computation.[23] Such a removal of both empirical conditions on the process and mathematical conditions on the processed properties is epitomized by Stevens' previously-quoted assertion that "measurement is the assignment of numerals to objects or events according to rule" (1959). On this basis, representational theories of measurement (Krantz et al., 1971) provided a mathematization (and in fact an axiomatization) of Stevens' standpoint, where the rule of assignment is required to be a morphism and no mathematical conditions are imposed on properties for their being measurable, and only very loose mathematical conditions are imposed on a process for being a measurement, as witnessed by the already quoted claim that "the theory of finite weak orderings [is] a theory of measurement, because of its numerical representation" (Suppes, 2002: p. 59):. Hence, this appears to be a case of position γ, as depicted in Fig. 4.5.

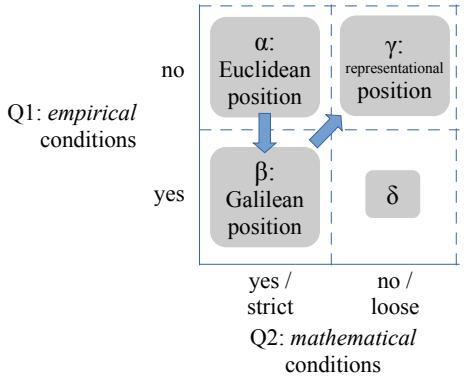


Fig. 4.5 The second transition: the representational position in the framework

4.4.2 Towards a different perspective?

The transition from the Galilean (β) position to the representational (γ) position provided a context for expanding measurability to non-physical properties, obtained at the price of removing empirical conditions on processes claimed to be measurements. This generality seems to explain why the representational theories are seldom used in physical sciences and engineering,[24] which remained stuck in the traditional, Galilean standpoint, as witnessed by the three editions of the VIM. As for empirical conditions, i.e., Q1, the first two editions (ISO, 1984; BIPM, 1993) defined measurement, rather implicitly, as a "set of operations having the object of determining the value of a quantity" (or "a value of a quantity", in the VIM2). A clearer position has been taken by the VIM3, which defines measurement as "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" (JCGM, 2012: 2.1) and then among the "presupposed" conditions lists "a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions" (JCGM, 2012: 2.1, Note 3). With respect to mathematical conditions, i.e.,

[23] See Sect. 6.5.1 for an analysis of this condition and of the critiques it has received.

[24] For a remarkable effort to adopt the representational approach in physical measurement, see several papers by Ludwik Finkelstein (e.g., 1984; 2003; 2005).

Q2, the concept <(measurable) quantity> has been redefined: while in the first two editions quantities were defined as properties with a measurement unit, i.e., properties representable on a ratio or an interval scale in Stevens' terminology, in the VIM3 the scope of measurement has been extended also to ordinal properties. On the other hand, according to the VIM3, "measurement does not apply to nominal properties" (JCGM, 2012: 2.1, Note 1), i.e., a mathematical constraint is still maintained on measurable properties, thus according to what could be considered a relaxed position β.

This reconstruction shows that all positions have been historically explored in the option space α–δ, with the exception of δ. Interestingly, it is exactly this position that we aim at better understanding – as depicted in Fig. 4.6 – in the chapters that follow, and that we eventually support.

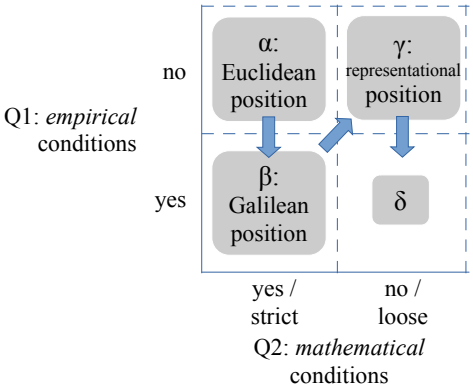


Fig. 4.6 The possible third transition in the framework

Prior to doing so, however, we conclude this chapter by discussing a philosophical position that we believe maintains the most valuable elements of each of the perspectives on measurement we have discussed so far without committing to their more problematic elements, and acknowledges the fundamental role of models in measurement, a stance we refer to as *model-dependent realism*.

4.5 A preliminary synthesis: model-dependent realism

There is a wide range of positions on measurement that could be described as realist in some sense, not all of which would commit to the idea that measurement is a process aimed at discovering the pre-existing – and therefore in some sense true – values of properties. Broadly, realist perspectives on measurement share some variant of the belief that at least one of the aims of measurement is to acquire information about properties, which are taken to in some sense exist independently of the measurement process, as well as the language, thoughts, and conventions of the individuals involved in the relevant measurement activities. Joel Michell (e.g., 2005) has argued that the "classical" understanding of the concept of measurement – the estimation of ratios of quantities, i.e., the measurand and the unit – entails realism about objects, properties, and numbers, the last being instantiated as ratios between individual quantities, in the Euclidean tradition. However, this is not the only possible version of a realist standpoint on measurement: for example, one can be realist about objects and properties without committing to realism about numbers, or agreeing with the assertion that measurement is always and only the discovery of such independently-existing ratios. Indeed, on sufficiently strict criteria such as those proposed by Michell (in reference to Otto Hölder's axioms, 1901), many well-accepted cases of physical measurement would fail to qualify as measurements: for example, electrical charge is not structured quantitatively in the sense given by Hölder, insofar as there exists a lower bound on possible electrical charge, i.e., the "elementary" charge (the charge of the

electron, according to the current theories), thus in violation of Hölder's second axiom (Mari et al., 2013).[25]

Realism is also sometimes misperceived as necessarily entailing a commitment to the possibility of absolute truth, or the position that there exists one true and complete description of the way the world really is. Particularly in the human sciences, this position is sometimes associated with an unflattering portrait of logical positivism or other forms of empiricism, and contrasted with nonrealist or antirealist views such as social constructionism or postmodernism (see Haig, 2014: Preface). In the context of the human sciences, the assumption that properties of objects must objectively exist in order to be measurable is often interpreted as implying a form of physical reductionism, and more specifically that the properties are exhaustively definable in biological (e.g., neurophysiological) terms; it is sometimes further concluded that there must exist a genetically-determined basis for variation in the property. Such claims may evoke a negative reaction from many scholars familiar with, for example, the controversial history of intelligence testing and its association with race and institutional racism (see Nisbett et al., 2012, for a recent review).

Of course, realism in general need not be associated with physical reductionism, and in fact the broad consensus amongst contemporary realist philosophers and philosophers of mind is that mental phenomena (properties, states, events, etc.) are no less a part of reality than physical phenomena (see in particular Dennett, 1991; Kim, 1998; Searle, 1992; see also Footnote 3 in Chapter 2). Perhaps even more importantly, realism need not be associated with a commitment to the possibility of absolute truth, nor do most contemporary philosophers formulate realist claims in this way. For example, Hilary Putnam (e.g., 1985; 1990; 1994), whose outlook was broadly realist throughout his career, argued that there are simply too many ways in which beliefs and symbols can be mapped onto the world for it to be plausible that there could be a single best description of the way the world is. In the context of the human sciences, it is often the case that there are several possible ways to describe psychological phenomena that are equally consistent with all available empirical data, and are nevertheless mutually inconsistent. For example, it can be shown that variation in personality characteristics as evaluated by a given set of responses to survey items is equally consistent with a model that describes personality as a typology (that is, in reference to a set of classes) and a model that describes personality in terms of continuous dimensions (e.g., Molenaar & Von Eye, 1994, cited in Borsboom, 2005).[26] As a separate but related point, it would seem to be a fairly straightforward observation that the meanings of terms about human beings, in both informal and formal discourse, are indexed to particular socio-historical conditions; for example, the meaning of a term such as "nursing competence" is likely to change over time (as new medical technologies are developed, social expectations and roles of hospital staff change, etc.), and from one geographical region to another, and even from one hospital ward to

[25] More generally, tThe fact that many physical properties are quantized would appear to lead to the paradoxical conclusion that they are not *really* quantities, and that they can be interpreted as quantities only in view of an approximate model that neglects the quantization. Hence, under Hölder's condition that only continuously varying quantities are measurable, the peculiar conclusion would be that all quantized properties are only *approximately* measurable. This has to do with the traditional distinction of quantities as either pluralities (or multitudes) or magnitudes, i.e., discretely or continuously divisible properties, thus based on a sharp distinction between counting ("how many") and measuring ("how much"). According to Aristotle, "'Quantum' means that which is divisible into two or more constituent parts of which each is by nature a 'one' and a 'this'. A quantum is a plurality if it is numerable, a magnitude if it is measurable. 'Plurality' means that which is divisible potentially into non-continuous parts, 'magnitude' that which is divisible into continuous parts." (Metaphysics, Book 5, Part 13; classics.mit.edu/Aristotle/metaphysics.5.v.html). The idea that measurability only refers to continuous quantities is desuete today.

[26] Examples of such pluralism are not limited only to the human sciences. A well known case from the physical sciences is about mechanical phenomena, for which Newtonian and relativistic mechanics are studied and operationally used, even though for some aspects they are incompatible (e.g., the speed of light in vacuum is relative in the former and constant in the latter). The justification of this multiplicity is pragmatic: at non-relativistic speeds the two theories basically provide the same results, and Newtonian mechanics is simpler than relativistic mechanics.

another (Maul, 2013). Similar comments could clearly be made about reading comprehension ability, as new kinds of texts and modes of interacting with texts are constantly being developed and introduced.

Scholars such as Jane Loevinger (1957) and Samuel Messick (1989) formulated versions of *constructive realism* that allow for the idea that properties measured by educational and psychological tests are, to an important extent, defined by socially-, culturally-, and historically-situated perspectives and concerns, as well as current theories of cognition, all of which may vary over time, and between different stakeholders at any given time.[27] An example of a philosophical framework that is consistent with such a view is found in Putnam's later (e.g., 2000) writings on *pragmatic realism* (also termed *natural realism*), which acknowledges that conceptual pluralism is not at odds with realism, but rather, "to use a Wittgensteinian idiom, *seeing* is always *seeing as*, and it is the interface between the world and the rich fabric of our concepts that jointly determines what we see" (Putnam, 2000: p. 20). On such an account, the existence of objective reality is not denied, but neither is it thought to be directly presented to our senses; instead, our conceptual and linguistic schemes and frameworks – both informally, and formally via *models* – actively shape our experiences and frame our knowledge of the world. Thus we organize and prioritize experience in a particular way, leading to the privileging of particular contrast classes, descriptive groupings, levels of explanation, and linguistic devices, and calling attention to specific observable facts, all of which might have been different for another observer or community of observers.

In the field of educational measurement, a sophisticated version of this position has recently been developed by Robert Mislevy (2018), who describes how within-person resources (e.g., developed forms of knowledge and skills) develop around between-person regularities in individuals' environments and experiences, then framing the challenge of human measurement in terms of assessing individuals' resources for engaging with high-level "linguistic, cultural, and substantive patterns" (p. 3) – which are real, albeit complex, features of the world – using data derived at the mid-level of personal experiences, which in turn are either approximately caused by or supervene upon low-level cognitive and neurophysiological processes.

This realism does not necessarily entail a naive correspondence view of truth that neglects the role of models in knowledge, nor does it deny that our conceptual schemes, models, and linguistic frameworks actively shape our experience of the world and frame our knowledge of it. However, in contrast to some of the more extreme formulations of relativism and conventionalism, it recognizes that models are always models *of something*, where in general the existence of the "something" does not depend[28] on the models.

An interpretation of such a pragmatic-realist view of measurement is illustrated in Fig. 4.7 below. According to the account implicit in this figure, the existence of natural reality is not denied, but neither is the fact that our various substantive and methodological theories and pragmatic concerns cause us to organize and prioritize experience in a particular way. This position is inspired by and

[27] As was briefly discussed in Sect. 4.3.2, one interpretation of the term "construct" as used in the human sciences is that it refers to a property that is in some sense constructed by us. As discussed by Earl Babbie (2013: p. 167), in reference to Kaplan's (1964) seminal analysis: "concepts such as compassion and prejudice are […] created from your conception of them, my conception of them, and the conceptions of all those who have ever used these terms. They cannot be observed directly or indirectly, because they don't exist. We made them up." (p. 168). As is argued in this section, however, it is fallacious to infer from the observation that we "made up" a property that its referents do not exist.

[28] Sometimes the very existence of modeled phenomena actually *does* depend to at least some extent on models, which then become illocutionary (Austin, 1975). For example, whether a set of neurophysiological facts about an individual make it more difficult for that individual to focus attention over long periods of time compared to other individuals is arguably a model-independent fact about that individual, but whether the individual has Attention Deficit Hyperactivity Disorder (ADHD) is a fact about how that individual has been labeled by other individuals, and is therefore at least partially dependent on a model of ADHD.

consistent with the Kantian metaphor that, when we look at the world, we do so not directly but through a particular lens, or set of lenses, which affect what we see. Further, although it may never be possible to look at reality without such lenses – and, as a corollary, it may never be possible to draw a hard line of demarcation between theory and observation, or theoretical and observational terms and concepts – we can still make considerable progress in understanding the limitations of our own understanding by acknowledging and examining our lenses as much as we are able, and where possible comparing them to the lenses of others. Fig. 4.7 visually illustrates this "lens" metaphor in the case of measurement: when we attempt to use measurement processes to look at some aspect of the world, what we see is jointly determined by the actual state of affairs and how we have chosen, explicitly or implicitly, to model (at least) the general property, the measurand, the environment, and the measurement process.[29]

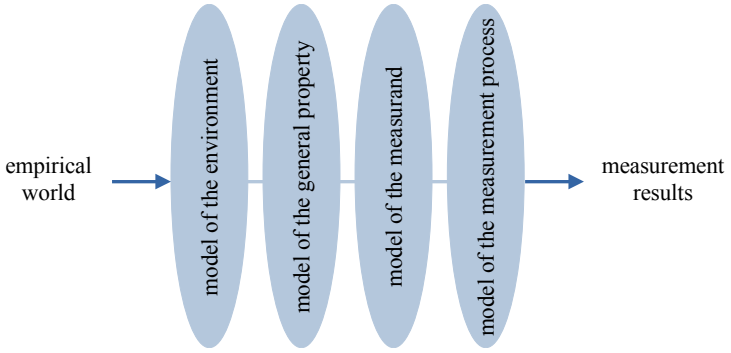


Fig. 4.7 A "lens" representation of the role of models in producing measurement results

Thus, properties could exist also only in consciousness, or at least in a societal context, and nevertheless be real and measurable. This point is perhaps most easily made with respect to psychosocial properties such as reading comprehension ability or well-being. In these cases, the definitions of the properties are indexed to particular sets of socio-historical conditions, and thus their definitions are at least in part delineated by contextually and pragmatically-driven frames of reference. Additionally, in the case of well-being (and other psychosocial properties such as desires, motivations, attitudes, and physical experiences such as pain and hunger, etc.), the existence of the property is partly or entirely dependent on the individual's experience of it. However, the subjective, first-person ontology of such properties does not preclude the possibility of others gaining knowledge of them in a publicly explainable, credibly documented way.[30]

Further, the connection between reality and a given measurement result is not in itself compromised by the fact that we choose to model the world in a particular way. In fact, models used in the service of scientific inquiry (including both substantive models, such as a cognitive-developmental model of learning in a particular domain, and psychometric models, such as the Rasch model) serve precisely the purpose of organizing experience. The acknowledgment that there is no single "true model" – that is, that models (and theories) are always underdetermined by facts – does not preclude the possibility of comparing models in terms of their quality. Thus, it is possible to maintain a realist perspective about the targets of measurement – objects and their properties – while acknowledging

[29] Fig. 4.7 is intended only as a rough visual representation of the role of models in producing measurement results, not as a representation of the whole measurement process, which (of course) involves actually performing measurements and as such entails more than simply looking at reality through the lens of models. The concepts used in the figure – <model of the general property>, etc. – are explained further in later chapters.

[30] Using terminology from Searle (1992), ontological subjectivity is not necessarily a barrier to epistemic objectivity (see also Maul, 2013). Or, using terminology from Dennett (1987), we may choose to model and study psychosocial properties by adopting a "design stance" or (especially) an "intentional stance" rather than a "physical stance".

that knowledge is constructed by humans, and can be so constructed in multiple ways. To make the point in a slightly different way, borrowing terms from Nancy Cartwright (1983) and Ian Hacking (1983), one can subscribe to *entity realism* without necessarily subscribing to *theory realism*; that is, the entities that feature in scientific theories may be regarded as real, without requiring a judgment about the truth of the theories into which they figure. On the other hand, conceptual pluralism is not the same as relativism: responsible science requires awareness and acknowledgment of the roles that conceptual frameworks, methodological approaches, and statistical models play in shaping investigations, and requires explication and empirical investigation of the hypothesized connections between the objects and processes under investigation and measurement results. Such awareness and acknowledgment is only possible to the extent to which claims are made explicit, and explication of claims requires a coherent semantics with which claims can be formulated.

In summary, a model-based realist perspective on measurement at the same time maintains something of each of the previous positions but rejects their most radical aspects:

- it accepts from realism the position that some properties do exist in the world, and are not just human constructs, but rejects the metaphysical claim that *values* of properties exist independently of our models (as illustrated in Fig. 4.7 by "empirical world" on the left-hand side);
- it accepts from non-realist empiricism (positivism, operationalism, representationalism, etc.) that empirical data can provide the evidential foundation for knowledge, but rejects the foundationalist claim that generic observation can have such a role, which is instead vested only in the empirical component of measurement systems, which are specifically designed for this purpose; moreover, it accepts that such evidence is always revisable (as illustrated in Fig. 4.7 by the "model of the measurement process" lens);
- it accepts from pragmatism (and relativism) that measurement is a designed-on-purpose process and that models in measurement are unavoidable, but rejects the possible conclusion that "anything goes" and that the quality of measurement results can be evaluated only *a posteriori*, in terms of the effectiveness of their application (as illustrated in Fig. 4.7 by the first 3 lenses on the left).

This perspective grounds the interpretation on an ontology and an epistemology of properties that is developed further in the chapters that follow.

References

American Educational Research Association (AERA) and other two Organizations (2014). *Standards for Psychological and Educational Tests*. Washington D.C.: American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME).

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford: Oxford University Press.

Babbie E. (2013). *The practice of social research* (13th ed). Belmont: Wadsworth.

Bell, S. (1999). *A beginner's guide to uncertainty of measurement*. Measurement Good Practice Guide No. 11 (2nd ed.). National Physical Laboratory. Retrieved from eprintspublications.npl.co.uk/1568

Bentley, J. P. (2005). *Principles of measurement systems*. New York: Pearson.

Bickhard, M. H. (2001). The tragedy of operationalism. *Theory and Psychology, 11(1)*, 35–44.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 36*, 35–37.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71(3)*, 425–440.

Borsboom, D. (2009). Educational measurement, 4th edition: Book review. *Structural equation modeling, 16*, 702–711.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*, 38–41.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.

Boumans, M. (2007). Invariance and calibration. In M. Boumans (Ed.), *Measurement in economics: A handbook* (pp. 231–248). London: Academic Press.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Campbell, N. R. (1920). *Physics: the elements*. Cambridge: Cambridge University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2)*, 81.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue internationale de philosophie*, 20–40.

Carnap, R. (1966). *Philosophical foundations of physics* (Vol. 966). New York: Basic Books.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Oxford University Press.

Chang, H. (2019). Operationalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/operationalism

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3,* 186–190.

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

De Morgan, A. (1836). *The connection of number and magnitude: At attempt to explain the fifth book of Euclid*. London: Taylor and Walton.

Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown and Company.

Deutsche Institut für Normung (DIN) (1995). *DIN 1319:1995. Fundamentals of metrology, Part 1: Basic terminology* (German standard). Deutsche Institut für Normung.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science, 1(1)*, 5–26.

Doignon, J. P. (1993). Geometry is measurement but measurement theory is more than geometry. *Journal of Mathematical Psychology, 37*, 472–476.

Edgeworth, F. Y. (1885). Observations and statistics: An essay on the theory of errors of observation and the first principles of statistics. *Transactions of the Cambridge Philosophical Society, 14*, 138–169.

Ente Italiano di Normazione (UNI) (1984). *UNI 4546:1984. Misure e misurazioni. Termini e definizioni fondamentali* (Italian standard). Milano: Ente Italiano di Normazione (in Italian).

Euclid's Elements of geometry, the Greek text of J.L. Heiberg (1883-1885) edited, and provided with a modern English translation, by Richard Fitzpatrick (2008). Retrieved from farside.ph.utexas.edu/Books/Euclid/Euclid.html

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., & Tucker, W. S. (1940). Final report of the committee appointed to consider and report upon the possibility of

quantitative estimates of sensory events. *Report of the British Association for the Advancement of Science, 2,* 331–349.

Finkelstein, L. (1984). A review of the fundamental concepts of measurement. *Measurement, 2*, 25–34

Finkelstein, L. (1994). Measurement: fundamental principles. In L. Finkelstein & K. Grattan (Eds.), *Concise encyclopedia of measurement and instrumentation* (pp. 201–205). Oxford: Pergamon.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement, 34*, 39–48.

Finkelstein, L. (2005). Problems of measurement in soft systems. *Measurement, 38*, 267–274.

Frank, P. G. (1956). *The validation of scientific theories.* Boston: Beacon Press.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*, 123–149.

Galilei, G. (1632). The Assayer. In S. Drake (Ed.), *Discoveries and Opinions of Galileo* (1957), New York: Doubleday.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan.

Gescheider, G. A. (2013). *Psychophysics: The fundamentals* (3rd ed.). Hoboken, NJ: Taylor and Francis.

Giordani, A., & Mari, L. (2012). Property evaluation types. *Measurement, 45*, 437–452.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory and Psychology, 2(3)*, 291–320.

Green, C. D. (2001). Operationalism again: What did Bridgman say? *Theory and Psychology, 11*, 45–51.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Hacking, I. (1990). *The taming of chance*. Cambridge: Cambridge University Press.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* Cambridge: MIT Press.

Heath, T. L. (1897). *The works of Archimedes*. Cambridge: Cambridge University Press.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass, Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, *Mathematisch-Physische Klasse, 53*, 1–46.

Hutton, C. (1795). *A mathematical and philosophical dictionary*. London: Johnson.

International Bureau of Weights and Measures (BIPM) and other six International Organizations (1993). *International Vocabulary of Basic and General Terms in Metrology (VIM)* (2nd ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML), International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP).

International Organization for Standardization (ISO) and other three International Organizations (1984). *International vocabulary of basic and general terms in metrology (VIM)* (1st ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML).

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.; 2008 version with

minor corrections). Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Kim, J. (1998). *Mind in a physical world*. Cambridge: MIT Press.

Koyré, A. (1948). *Du monde de l'à peu près à l'univers de la précision*. In A. Koyré (Ed.), Etudes d'histoire de la pensée philosophique (pp. 341-362). Paris: Gallimard.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971, 1989, 1990). *Foundations of measurement – Additive and Polynomial Representations* (Vol. I). New York: Academic Press.

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

Kyburg, H. E. (1984). *Theory and measurement*. Cambridge University Press.

Leahey, T. H. (1980). The myth of operationism. *The Journal of Mind and Behavior, 1(2)*, 127–144.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Margenau, H. (1958). Philosophical problems concerning the meaning of measurement in physics. *Philosophy of Science, 25*, 23–33.

Mari, L. (1997). The role of determination and assignment in measurement. *Measurement, 21*, 79–90.

Mari, L. (2003). Epistemology of measurement. *Measurement, 34*, 17–30.

Mari, L. (2013). A quest for the definition of measurement. *Measurement, 46*, 2889–2895.

Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: A pragmatic view. *IEEE Transactions on Instrumentation and Measurement, 61*, 2107–2115.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

Maul, A. (2013). The ontology of psychological attributes. *Theory & Psychology, 23*, 752–769.

Maul, A. (2014). Justification is not truth, and testing is not measurement: Understanding the purpose and value of the Standards. *Educational Measurement: Issues and Practice, 33*, 39–41.

Maul, A. (2018). Validity. In B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1771–1775). Thousand Oaks: SAGE Publications.

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311–320.

McGrane, J. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology, 6*, 431.

McGrane, J., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement, 152*. doi.org/10.1016/j.measurement.2019.107346

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillian.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* New York, NY: Psychology Press.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology, 14*, 121.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement, 38(4)*, 285–294.

Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 135–170). Charlotte, NC: Information Age Publishing.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge.

Molenaar, P. C. M., & Von Eye, A. (1994). On the arbitrary nature of latent variables. In A. Von Eye & C. C. Clegg (Eds.), *Latent variables analysis*. Thousand Oaks: Sage.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23(2)*, 5-12.

Narens, L. (1985). *Abstract measurement theory*. Cambridge, MA: MIT Press.

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Lawrence Erlbaum.

Narens, L., & Luce, R. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99(2)*, 166–180.

Neurath, O., Hahn, H., & Carnap, R. (1973). *The scientific conception of the world: The Vienna Circle. Empiricism and sociology*. Boston: Reidel.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspectives, 10(1–2)*, 1–29.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New Findings and Theoretical Developments. *American Psychologist, 67*, 130–159.

Putnam, H. (1985). *Realism and reason (volume 3 of Philosophical Papers).* Cambridge: Cambridge University Press.

Putnam, H. (1990). *Realism with a human face*. Cambridge, MA: Harvard University Press.

Putnam, H. (1994). *Words and life*. Cambridge, MA: Harvard University Press.

Putnam, H. (2000). *The threefold cord: Mind, body and world*. New York City: Columbia University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Regtien, P. P. L. (2004). *Measurement science for engineers*. London: Elsevier.

Rossi, G. B. (2006). An attempt to interpret some problems in measurement science on the basis of Kuhn's theory of paradigms. *Measurement, 39*, 512–521.

Rossi, G. B. (2007). Measurability. *Measurement, 40*, 545–562.

Rossi, G. B. (2014). *Measurement and probability: A probabilistic theory of measurement with applications*. Dordrecht: Springer.

Rossi, G. B., & Crenna, F. (2013). On ratio scales. *Measurement, 46(8)*, 2913–2920.

Russell, B. (1903). *The principles of mathematics*. Cambridge: Cambridge University Press.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Shepard, L. A. (1993). Evaluating Test Validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 19). Washington, DC: AERA.

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270.

Skinner, B. F. (1971). *Beyond freedom and dignity*. New York: Knopf.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Cham: Springer.

Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 13*, 4–12.

Speitel, K. (1992). Measurement assurance. In G. Salvendy (Ed.), *Handbook of Industrial Engineering* (pp. 2235–2251). New York: Wiley.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In C. West Churchman & P. Ratoosh (Eds.), *Measurement, definitions and theories* (pp.18–63). New York: Wiley.

Stevens, S. S. (1975). *Psychophysics: introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education, 101*, 60–70.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Suppes, P. (2002). *Representation and invariance of scientific structures*. CSLI Publications.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. *Handbook of mathematical psychology, 1*, 1–76.

Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/measurement-science

Teller, P. (2013) The concept of measurement-precision. *Synthese, 190*, 189–202.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47(1)*, 65–72.

Weyl, H. (1949/2009). *Philosophy of mathematics and natural science*. Princeton, NJ: Princeton University Press.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46(9)*, 3766–3774.

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice, 37(1)*, 5–20.

Chapter 5.
What is measured?

This chapter aims to explore some key components of an ontology and an epistemology of properties. What is evaluated, and more specifically, measured, are properties of objects, such as lengths of rigid bodies and reading comprehension abilities of individuals, and the results of evaluations, and thus of measurements, are values of properties. Hence, a study of the nature of properties and of our ways of securing knowledge of them is a pivotal component of measurement science. We start from the hypothesis that properties of objects are associated with modes of empirical interaction of the objects with their environment. Consistently with the model-dependent realism introduced in Chapter 4, the Basic Evaluation Equation

property of a given object = value of a property

of which the relation

measurand = measured value of a property

is a specific case, is interpreted as a claim of an actual referential equality, which conveys information on the measurand because the measurand and the measured value remain conceptually distinct entities.

5.1 Introduction

Measurement is a specific kind of evaluation of empirical properties of objects. A measurement-oriented ontology and epistemology of properties is a complex subject, in which the ontic dimension (*what properties are*) and the epistemic dimension (*what we can know about properties and how we can know it*) are deeply intertwined. Hence, a discussion of properties is an important part of a discourse on measurement. Let us restart from our discussion about properties in Sect. 2.2.

An empirical property of an object is associated with *a mode of empirical interaction of the object with its environment*, where this association is understood such that

- *an object empirically interacts with its environment in multiple modes, and each mode of interaction is considered to correspond to a property of the object,*[1] *and*
- *some objects are comparable with respect to some of their properties, and sometimes distinct objects are discovered to have empirically indistinguishable properties.*[2]

By considering properties of objects as associated with modes of empirical interaction of objects with their environment, we avoid taking a position on

[1] This does not preclude that in some cases distinct modes of interaction correspond to the same property. A well-known example in physics is mass, which is associated with both inertial and gravitational phenomena, and therefore to distinct modes of interaction. This is consistent with Brian Ellis' insight that "our quantity concepts are generally cluster concepts" (1968: p. 32). The interaction of an object with different instruments might be interpreted as different modes of interaction, and only from a radically operationalist perspective the interaction with each instrument would correspond, by definition, to a different property (see Sect. 4.2.2).

[2] The plausibility of these conditions is confirmed also in a philosophical context. For example, to the question *What is a property?* Baron et al. answer: "This is a thorny issue. For present purposes we conceive of properties as the entities that ground causal powers and similarity relations." (2013: p. 35).

- the nature of properties of objects, other than that we consider them to be entities able to produce at least in principle observable effects and to support the comparison of objects,[3]
- the difference between inherent (or essential) and contingent (or accidental) properties, where the identity of the bearer, i.e., the object, is supposed to be affected by a change of its inherent properties (a vase might be considered a different object than the amorphous amount of clay from which it was shaped), but not by a change of its contingent properties (a vase may still be considered the same object, even if it is chipped), and
- the issue of the possible distinction between object-specific ("primary", e.g., mass) and observer-related ("secondary", e.g., perceived color) properties.[4]

This is consistent with a pragmatic stance which, we believe, is appropriate to ground a measurement-oriented discussion on properties:

- we consider a property of an object to *exist* insofar as the object somehow *interacts* with its environment, though we accept that everything we consider as known about a property is always revisable, and could turn out to be wrong;
- we acknowledge that empirical interactions may be physical, but also psychological, sociological, etc., thus admitting the existence of non-physical properties;
- we consider a property of an object to be associated with a mode of interaction of the object, but we acknowledge that the existence of a property may be hypothesized also independently of a mode of interaction, and we do not say anything further about what a property *is per se*.[5]

Furthermore, we consider that phrases such as "the objects *a* and *b* are comparable with respect to a given property" and "a given property of the object *a* and a given property of the object *b* are comparable" refer to the same empirical situation.[6] Whenever this happens, we say that the given properties of *a* and *b* are *of the same kind* (as in JCGM, 2012: 1.2), and thus we assume that a *general property* exists of which the given properties of *a* and *b*, which we call *individual properties*, are

[3] That "the familiar objects of the everyday world agree in their characteristics, features, or attributes" is considered "a prephilosophical truism" (Loux & Crisp, 2017: p. 18).

[4] For an analysis on this distinction, see for example Heil (2003: chapter 8), who also presents the idea that properties manifest themselves through empirical interactions of objects: "all there really is to a concrete entity is its power to affect and be affected by other entities. Assuming that an entity's powers depend on its properties, this suggests that there is no more to a property than powers or dispositionalities it confers on its possessors. [...] The business of science is to tease out fundamental properties of objects. Properties are what figure in laws of nature, and laws govern the behaviour of objects. Properties, then, are features of the world that make a difference in how objects behave or would behave." (p. 75).

[5] The subject of *what properties are* is complex, and out of the scope of this book, in which a (black box) characterization of *how properties manifest themselves* is sufficient. For a philosophically oriented introduction to this subject, see for example the articles by Orilia and Swoyer (2020), Weatherson and Marshall (2018), and Wilson (2017) in the Stanford Encyclopedia of Philosophy. Concepts such as <property>, <attribute>, <feature>, <characteristic>, etc. are so fundamental in human knowledge that it is not clear how they could be defined without circularity. For example, René Dybkaer (2004: p. 51) defines <property> as an "inherent state- or process-descriptive feature of a system including any pertinent components", while leaving <feature> as a primitive (i.e., undefined) concept. Were he requested to define <feature>, he plausibly might have included the term "property" in the definition, thus showing that the concept is ultimately defined in terms of itself. Not surprisingly, these concepts are also sometimes used in a somewhat confusing way in pivotal texts of measurement science, for example *Foundations of Measurement* (Krantz et al., 1971: p. 1) which states in its opening sentence: "when measuring some attribute of a class of objects or events, we associate numbers (or other familiar mathematical entities, such as vectors) with the objects in such a way that the properties of the attribute are faithfully represented as numerical properties". The idea that an attribute has properties represented as properties is not exactly obvious, to say the least.

[6] In fact, the first phrase, while seemingly better reporting what empirically happens ("in the comparison we handle objects, right?", as a colleague of ours told us), assumes a greater ontic burden, given that the entity with respect to which objects are compared is a kind of property, whereas kinds of properties do not explicitly appear in the second phrase. Furthermore, it is surely possible that properties of the same object are compared, for example the height and the width of a rigid body: in terms of comparison of objects this would require a cumbersome phrasing like "an object is compared with itself with respect to a (kind of) property".

instances (see also an introduction of these concepts in Sect. 2.2). General and individual properties – such as mass and any given mass, respectively – are sometimes called, particularly in the philosophical literature, "determinables" and "determinates", respectively; see, e.g., Wilson, 2017). Hence, length and reading comprehension ability are examples of general properties, and the length of a given rod and the reading comprehension ability of a given individual are examples of individual properties. The length of a given rod and the wavelength of a given radiation are comparable, being individual properties of the same kind, i.e., the general property length, whereas the length and the mass of any two objects are not comparable, being individual properties of different kinds.

Even from this philosophically modest perspective, several important issues remain open for consideration, regarding in particular the distinction between the *existence* of a property and the *knowledge* that we can have of it. This chapter is devoted to an analysis of this subject, and to providing an interpretation of the (measurement) relation

measurand = measured value of a property

as introduced in Sect. 2.2.4, a specific case[7] of the *Basic Evaluation Equation*

property of a given object = value of a property

formalized as[8]

$P[a] = p$

for example

length[*rod a*] = 1.2345 m

or

reading comprehension ability[*student b*] = 1.23 logits (on a specific RCA scale)

or

blood type[*patient c*] = A in the ABO system

where in the first case the relation is about a ratio (and more specifically an additive) quantity (length) and the value is reported as the product of a number and a quantity unit (the metre), in the second case the relation is about an interval quantity (reading comprehension ability) and the value is reported as a number in an interval scale (here denoted as logits, on a specific RCA scale) (Maul et al., 2019), and in the third case the relation is about a nominal property (blood type) and the value is reported as an identifier for a class in a specified classification system (the ABO system) (Mari, 2017).

In the case of ratio quantities, which is the common situation in the measurement of physical properties, the relation becomes

quantity of a given object = value of a quantity

formalized more specifically as

$Q[a] = \{Q\}[Q]$

where the value is the product of a number $\{Q\}$ and a quantity unit $[Q]$ (which is a different usage of "[ ]" than on the left-hand side of the equation): thus, in the example above, $\{$*length*[*rod a*]$\}$ = 1.2345 and [*length*[*rod a*]] = m.[9] While most of what follows applies to all properties, independently of their type, we often refer more specifically to quantities, due to their widespread use in the tradition of

[7] Measurands are "quantities intended to be measured" (as defined in JCGM, 2012: 2.3): hence, while measurands are properties of objects, a property of an object becomes a measurand for us only when we are interested in obtaining a value for it via a measurement. Several aspects of our analysis apply to the generic case, thus for example also to Basic Evaluation Equations which describe specifications, instead of reporting results of measurements.

[8] As noted in Sect. 2.2.3, the notation $P[a]$ is aimed at highlighting that P can be formalized as a function, but it is not a mathematical entity as such.

measurement science and their richer algebraic structure, which makes examples easier to present and understand.

Box 5.1 – A very short introduction to ontology

Given our statement that this chapter is mainly devoted to exploring a measurement-oriented ontology of properties, a few preliminary words might be useful about what we consider to be an ontology. As Willard V. O. Quine wrote, "A curious thing about the ontological problem is its simplicity. It can be put in three Anglo-Saxon monosyllables: 'What is there?' It can be answered, moreover, in a word – 'Everything' – and everyone will accept this answer as true. However, this is merely to say that there is what there is. There remains room for disagreement over cases; and so the issue has stayed alive down the centuries." (1948: p. 21).

Cases about which there can be or has been disagreement include the sphere of fixed stars, phlogiston, and continuous flows of electricity, not to mention nearly every proposed property in the human sciences, perhaps most famously general intelligence: it was not at all trivial to arrive at the conclusion that, for example, phlogiston does not exist, and in fact this required a radical revision of several related bodies of knowledge. But of course even today we can talk in a meaningful way about the sphere of fixed stars, phlogiston, and continuous flows of electricity; otherwise a sentence such as "all visible stars are fixed to a celestial sphere" would be meaningless rather than false (on par with a phrase like "all qwerty uiop are fixed to a celestial sphere").

Hence, the fact that x appears in meaningful sentences is not sufficient to conclude that x exists. It should be acknowledged that there are indeed different *modes of existence*: for example, both paper books and prime numbers greater than 1 million exist, but their modes of existence differ. The "disagreement over cases" to which Quine refers is related to the existence in a given mode, not to the generic situation of any possible mode of existence. As another well-known example, unicorns do not exist as physical entities, but do exist as literary entities. Thus, our ability to talk in a grammatically correct way of an entity x is not sufficient to guarantee that x exists as an entity of the kind Y, nor is the fact that x exists as an entity of the kind Z is sufficient to guarantee that it exists also as an entity of the kind Y. As a consequence, a claim of existence of an entity x is interesting only if it is specified as a claim of Y-existence, i.e., existence in the mode Y, for a given Y. The ambiguity is avoided if instead of the generic "do unicorns exist?" we ask "do unicorns exist as physical entities?", i.e., "do unicorns have Y-existence?", where Y = physical, which according to our current knowledge has a different answer from "do unicorns have Z-existence?", where Z = literary.

Furthermore, it is important to recognize that such a claim is about the Y-existence of x as an object, not about the meaning of the term "x", or about the existence of the concept <x>. As it was put by Quine: "The phrase 'Evening Star' names a certain large physical object of spherical form, which is hurtling through space some scores of millions of miles from here. The phrase 'Morning Star'

[9] As inspired by the seminal work of James Clerk Maxwell (1873), this relation is commonly written as $Q = \{Q\}\,[Q]$

which we call "Q-notation" for short. Despite its success (see, e.g., de Boer, 1995: p. 405 and Emerson, 2008: p. 134, but also JCGM, 2012: 1.20 Note 2 and ISO, 2009b), this notation is not completely clear, as it does not maintain the distinction between general quantities and individual quantities. By writing the left hand side entity as $Q[a]$ we make explicit the reference to the quantity Q of the object a. This also highlights that, while the unit is a feature of the general quantity Q, and therefore writing it as $[Q]$ is correct (recalling that brackets in "$Q[a]$" and "$[Q]$" have different meanings: "$Q[a]$" stands for the Q of a, e.g., the length of a given rod; "$[Q]$" stands for the unit of Q), the numerical value depends on both the quantity of the object and the chosen unit, and as such it should be more correctly indicated as $\{Q[a]\}_{[Q]}$, to be read "the numerical value of $Q[a]$ in the unit $[Q]$" (see also ISO, 2009b: 6.1).

names the same thing, as was probably first established by some observant Babylonian. But the two phrases cannot be regarded as having the same meaning; otherwise that Babylonian could have dispensed with his observations and contented himself with reflecting on the meanings of his words. The meanings, then, being different from one another, must be other than the named object, which is one and the same in both cases." (1948, p.28). In Gottlob Frege's terminology (1892), this is effectively presented by acknowledging that two terms (such as "Evening Star" and "Morning Star") can have different *senses* and nevertheless the same *reference*, and also that a term can have *Y*-sense but no *Y*-reference, i.e., a term can be intended as referring to an entity *x* of the kind *Y* even though *x* has no *Y*-existence (for example, we may well understand the claim that some unicorns exist as biological entities, and nevertheless – or, in fact, exactly because we understand exactly its meaning – consider it false according to the best currently available knowledge). Furthermore, this explains the difference between the two relations

Evening Star = Morning Star

and

Evening Star = Evening Star

The former required a lot of astronomical ingenuity and knowledge for its discovery, while the latter is a trivial, logical truth, as is any identity $x = x$, independent of astronomical facts.

It is then worth emphasizing that, as used here, "ontology" is not a synonym of "metaphysics". Rather, "it refers to the set of 'things' a person believes to exist, or the set of things defined by, or assumed by, some theory. What's in your ontology? Do you believe in ghosts? Then ghosts are in your ontology, along with tables and chairs and songs and vacations, and snow, and all the rest." (Dennett, 2017: p. 60) (for a wide presentation of a "scientific perspective" on ontology, see Bunge, 1977, possibly starting from his "list of ontological principles occurring in scientific research", p. 16). Of course, at least some of the things we believe to exist are physical, and some are psychosocial.

A key problem in ontology – perhaps *the key problem* of any ontology of properties – is whether things such as a given mass and mass as such are existing (though possibly abstract) entities, or are just concepts that we produce for organizing our knowledge. An intermediate position, called *extensionalism*, is that they are not entities as such but sets (or possibly mereological sums: Varzi, 2019) of them: a given mass would then just be a set of masses of objects, and mass the set of all given masses, and therefore a set of sets.

The answer to such a problem depends on whether one's ontology has room for abstract entities or only for concrete entities, a distinction that is sometimes presented in terms of *universals* and *particulars* (the possible differences between <abstract> and <universal> and between <concrete> and <particular> are beyond the scope of our purposes here), and grounds the opposition between *realism* and *nominalism*: "The realist's ontology represents a two-category ontology; it postulates entities of two irreducibly different types: particulars and universals. According to the nominalist, however, all the theoretical work done by the two-category ontology of the realist can be done by an ontological theory that commits us to the existence of entities of just one category, particulars." (Loux & Crisp, 2017: p. 50).

In what follows we try to remain as neutral as possible about the alternative between realism and nominalism, and mention the position that we believe is more appropriate for effectively accounting for the key facts of measurement results only in Sect. 5.3.

5.1.1 The possible meanings of the Basic Evaluation Equation

The Basic Evaluation Equation

property of a given object = value of a property

conveys the core information obtained by measurement (neglecting measurement uncertainty, for the moment). Despite the fact that information of this sort is commonly produced and used, the apparent simplicity of the relation hides the question: *is this relation an actual equality, or is the "=" sign in it just a placeholder for a different relation?*

The problem is mostly immaterial in day-to-day practice and is thus usually left in the background, so that one sometimes encounters claims such as Gary Price's (2001: p. 294) that the relation "'equals' means 'is expressed, modeled, or represented by'". Since <expression>, <modeling>, and <representation> are distinct concepts, and none of them is the same as <equality>, it seems that such a statement only informs us of a lack of interest in understanding what kind of information a Basic Evaluation Equation actually conveys. In distinction to this vagueness, it is our position that an answer to this problem is indeed critical for a measurement-related ontology and epistemology of properties. Note that different positions are possible (Mari, 1997), the two extremes being

- a strong ontology, which assumes that properties of objects inherently have values, so that if they are known it is because they have been *discovered* by means of experimental activities, and
- a weak ontology, which assumes that values are *assigned* as a means of representation.

The common ground of these positions is the acknowledgment that (i) in performing measurement, the starting point is the identification of a property of an object to be measured, i.e., the measurand, and the discovery (according to a strong ontology) or the selection (according to a weak ontology) of a set of possible values of that property, and that (ii) the outcome of the process is that one value in the set (or a subset of them, as in JCGM, 2012: 2.1, if measurement uncertainty is taken into account) is attributed to the measurand. The issue is about how to interpret such an attribution: is the value established because it exists in the object before and independently of any experimental activity, or is it (just) chosen to suitably report the acquired information on the object? Is then measurement akin to *discovery* or *invention*? Even more fundamentally, this issue is grounded upon the issue of the very existence of properties: do entities such as length and reading comprehension ability exist in the world, or are they just constructs by which we organize our knowledge?

Positions like the one underlying the representational theories of measurement (see Sect. 4.2.3) emphasize the representational aspect of measurement, plainly stating that the task of measurement is "to construct numerical representations of qualitative structures" (Krantz et al., 1971: p. xviii), and from their beginnings have acknowledged that "the major source of difficulty in providing an adequate theory of measurement is to construct relations which have an exact and reasonable numerical interpretation" (Scott & Suppes, 1958: p. 113). Such positions are plausibly based on weaker, less demanding ontologies, and this may make their practical consequences applicable also to those who accept a stricter position: a value *may be chosen to represent* a property of an object exactly because that property *has* that value.[10] If it is maintained that properties of objects do not inherently have values, the representation may be chosen according to different criteria, and is required to be at least consistent: if properties of objects are observed to be ordered then their assigned values should be ordered in turn, but any ordered set would be suitable to perform such a purely symbolic task, and so

[10] Interestingly, the form of the Basic Evaluation Equation, in which the property of the object and the value of the property are related by an "=" sign, suggests the interpretation that the property *is* the value: we discuss this delicate point in Sect. 6.4.

on. However, a stronger ontology invites interpretation of advancements in measurement-related knowledge and practices as an evolutionary process: at the beginning the available information could be so "meager and unsatisfactory" (quoting Lord Kelvin; see the related discussion in Kuhn, 1961) that the evaluation results are more or less everything that is known of the considered property, and therefore consistency in the representation is the only condition that can be sought. Such an approach could be later abandoned with the acquisition of more and better information, leading to corroboration of the hypothesis of the very existence of the property, up to the extreme position that the measurand has a knowledge-independent *true value*, to be estimated through measurement.[11]

5.1.2 A pragmatic introduction to the problem

Basic Evaluation Equations are at the core of any measurement, and therefore an understanding of them is a requirement for a well-grounded measurement science. However, "equality gives rise to challenging questions which are not altogether easy to answer" (Frege, 1892: p. 25): quoting again Price (2001: p. 294), in a Basic Evaluation Equation does the equality sign mean "is expressed, modeled, or represented by", or (in some sense to be specified) equality, or something else?

In what follows, we introduce the problem only in terms of ordinal comparisons, thus about what could be called a "Basic Evaluation *Inequality*": this property of this object is less than this value. Our claim is that this does not affect the generality of the presentation, and hopefully makes it clearer by referring to a less controversial relation than equality, thus avoiding the "challenging questions" to which Frege alluded – including the ones connected with the possible role of uncertainty – that equality brings.

Let us consider the case of mass.

1. In measurement we deal with entities such as masses of given objects, e.g., *mass*[*rod a*], and values of mass, e.g., 1.23 kg. For the sake of argument, let us call the former "O-entities" (i.e., related to objects) and the latter "V-entities" (i.e., related to values), by noting that different terms may (but do not necessarily) correspond to different kinds of entities, and the conclusion that we might well reach is that properties of objects, i.e., O-entities, and values of properties, i.e., V-entities, are different ways of referring to the same kind of entity (O-entities and V-entities are more specifically termed "addressed quantities" and "classifier quantities" respectively by Mari and Giordani, 2012).
2. O-entities and V-entities are such that we can compare
   - O-entities among themselves (the mass of rod *a* is less than the mass of rod *b*),
   - V-entities among themselves (1.23 kg is less than 2.34 kg), and
   - O-entities and V-entities (the mass of rod *a* is less than 1.23 kg).
3. In particular, the chain of inequalities

   $mass[rod\ a] < 1.23\ kg < 2.34\ kg < mass[rod\ b]$

[11] The history of the concept of true value is complex, and definitely still not settled (see also Sects. 3.2.2, 4.2.1, and 7.4). While sometimes considered to be a useless metaphysical residual, in some contexts the reference to true values is maintained and emphasized. As an example, the current version of the *NIST Quality Manual for Measurement Services* (the National Institute of Standards and Technology, NIST, is the US National Metrology Institute) defines <measurement> as (emphasis added): "an experimental or computational process that, by comparison with a standard, produces *an estimate of the true value* of a property of a material or virtual object or collection of objects, or of a process, event, or series of events, together with an evaluation of the uncertainty associated with that estimate, and intended for use in support of decision-making." (11th version, 2019, www.nist.gov/system/files/documents/2019/04/09/nist_qm-i-v11_controlled_and_signed.pdf, including the note that "the NIST Measurement Services Council approved [this] definition of measurement to include value assignments of properties using qualitative techniques") (Possolo, 2015: p. 12).

is understandable and does not pose problems of interpretation, also about the transitivity of the relation (e.g., from *mass*[*rod a*] < 1.23 kg and 1.23 kg < 2.34 kg the conclusion is unproblematically obtained that *mass*[*rod a*] < 2.34 kg), thus justifying the hypothesis that, at least at a sufficiently abstract level, what is designated here by "<" is actually the same relation applied to both O-entities and V-entities.

4. The comparison of O-entities among themselves and the comparison of V-entities among themselves are different processes:
   - for comparing O-entities among themselves we compare properties of objects, thus by means of an empirical process, such as the one performed by means of a balance and leading to the possible conclusion that the mass of rod *a* is less than the mass of rod *b*;
   - for comparing V-entities among themselves we compare numbers (assuming their unit is the same), thus by means of a mathematical process, such as the one that leads to the analytical conclusion that 1.23 kg is less than 2.34 kg.
5. A correspondence can be established between O-entities and V-entities:
   - O-entities can be made to correspond to V-entities through a process that can be generically called "evaluation" (see Chap. 2, and particularly Sect. 2.2.4, for discussion of this lexical choice), leading to the association of a value with the given property of an object; measurement is then a specific kind of evaluation: the mass of rod *a* can be evaluated as 0.12 kg;
   - V-entities can be made to correspond to O-entities through a process that can be generically called "realization", leading to the selection or construction of an object such that one of its properties is associated with the given value: 0.12 kg can be realized by the mass of rod *a*.
6. Via these correspondences, if O-entities are evaluated, then the obtained V-entities can be compared among themselves, and this comparison conveys information about the relevant O-entities, so that the O-entities do not themselves need to be directly compared (as in stage (3) above): if the mass of rod *a* is evaluated as 0.12 kg and the mass of rod *b* is evaluated as 3.45 kg, the comparison that 0.12 kg is less than 3.45 kg leads to the inference that the mass of rod *a* is less than the mass of rod *b* where such an inference is valid if the two evaluations are valid in turn. Nevertheless, comparing O-entities among themselves and comparing V-entities among themselves remain different processes: the former is an empirical process, the latter is an analytical process.
7. Hence, O-entities and V-entities are at the same time
   - *analogous in some respects*, because they are comparable: both O-entities and V-entities can be thought of as properties, but
   - *different in some other respects*, because the ways in which we compare them among themselves are different: O-entities are properties identified through objects that have them, whereas V-entities are properties identified through numbers that multiply units.[12]

Fig. 5.1 summarizes the relations among these entities.

[12] We have not been able to find a term to designate the entities obtained by multiplying or dividing a unit by a number which is not necessarily integer (a multiple of a unit is a "measurement unit obtained by multiplying a given measurement unit by an *integer* greater than one" (JCGM 2012: 1.17), and a submultiple of a unit is a "measurement unit obtained by dividing a given measurement unit by an *integer* greater than one" (JCGM 2012: 1.18, emphasis added). Hence, we maintain the term "multiple" with this broader meaning: accordingly, if u is a unit and *x* is a non-negative real number, *x* u is a multiple of u.

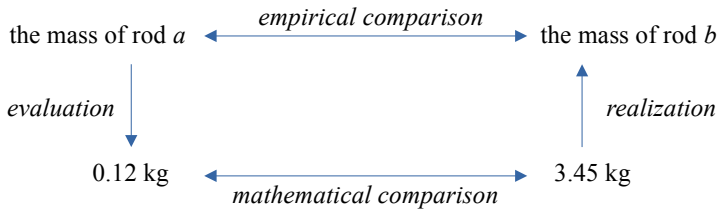


Fig. 5.1 Graphical representation of the relations among object-related entities, such as the mass of some given object, and value-related entities, such x kg for some given positive number x

What follows in this and the following chapter is an exploration and an analysis of these fundamental issues.

5.1.3 Anticipating the main outcomes

As we have already seen, a measurement-oriented ontology and epistemology of properties is definitely a non-trivial subject. To help follow and orient the analysis that follows, we start by anticipating here some of the main conclusions.

For any given property, say mass, there are four interrelated but conceptually distinct kinds of entities that can be taken into account:

(i) the general property (e.g, mass), M;

(ii) individual properties (e.g, given masses), m;

(iii) properties of given objects (e.g, the masses of given objects a), $M[a]$;

(iv) values of the property (e.g, x kg for any given appropriate x): 1.2345 kg.

Our basic claim is that all four of these kinds of entities are required in a sufficiently well structured discourse on measurement.

(i) General properties are the entities that measuring instruments are designed to measure, so that for example balances are designed to measure masses, not the mass of any given object in particular; moreover, scales (and therefore, in particular, units) are about general properties. Scientific laws, when they are invoked, pertain first to general properties, and the same holds for dimensional analysis (for example, when stating that density has the dimension of mass times length to the minus three we refer to density, mass, and length as such, and not to any given density, any given mass, and any given length).

(ii) Individual properties are the entities whose relations characterize the mathematical structure of the general property of which they are instances: mass is an additive quantity because the set of masses, independently of the objects that can have such masses and their relation with any possible unit of mass, has an additive structure (for example, when stating that for any two positive masses their addition / composition / concatenation is greater than either of them we may refer to individual masses as such, and not necessarily to the masses of given objects or to some values of mass).

(iii) Properties of given objects are the entities that are measured in any actual measurement and to which values of properties are attributed: a given balance in a given situation is an instrument for measuring the mass of a given object.

(iv) Finally, values of properties are the entities that report the information acquired by means of calibrated measuring instruments applied to properties of objects: 1.2345 kg and 2.7216 lb are values that could be attributed to the mass of a given object.

While all this could well be taken for granted, the structure of the relations among these entities is important for a measurement-oriented ontology and epistemology of properties. Starting from the

uncontroversial assumption that *individual properties* (ii)[13] *are instances of general properties* (i) (so that for example, any given mass is an instance – i.e., an example, a case, … – of mass, i.e., more customarily and trivially, any given mass is a mass), we acknowledge that individual properties are entities which need to be somehow identified in order to be handled, and in particular to be compared with each other. On this basis we develop here two basic arguments.

First, *individual properties* (ii) *are identified as properties of given objects* (iii) *or as values of properties* (iv), so that for example a given mass can be identified as the mass of a given object *a* or as *x* kg for a given non-negative number *x*, as it is explicit in the case of the Basic Evaluation Equation. In fact, properties of objects and values of properties are complementary modes of identification of instances of general properties: by identifying a mass as the mass of a given object *a*, the reference is to the object *a* that bears that mass; if a mass is instead identified as *x* kg for a given non-negative number *x*, the reference is to an element of the structure that, via the choice of a mass unit and the construction of its multiples, includes all masses.[14]

Fig. 5.2 summarizes the relations among these kinds of entities, and highlights the pivotal role of individual properties in the conceptual framework we are developing. An individual property p is an instance of a general property *P*, and can be identified as the property *P*[*a*] of an object *a* or as the value *p* of property *P*.

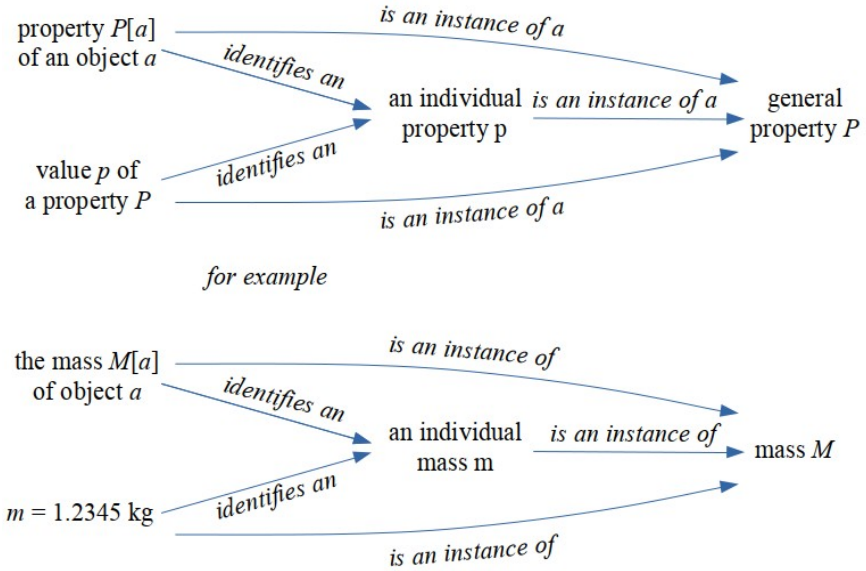


Fig. 5.2 Graphical representation of the relations among the four kinds of entities related to properties: the generic model (top) and an example (bottom)

Second, the complementarity of the two modes of identification of individual properties is exploited in measurement. The information conveyed by a Basic Evaluation Equation

property of a given object = value of a property

is indeed that an individual property p_1, identified as a property of a given object (i.e., $p_1 = P[a]$ for a given object *a*), and an individual property p_2, identified as a value of a property (i.e., $p_2 = p$ for some

[13] The labels "(i)", "(ii)", etc. refer here to the four assertions immediately above.

[14] It is not controversial that individual properties are instances of general properties, but there is a delicate ontological issue about how properties of objects and values of properties are related to individual properties (see also the discussion in Sect. 5.3.1). As written, with the aim of remaining as independent as possible of ontological presuppositions, we only assume that properties of objects and values of properties identify individual properties, where this relation may be alternatively characterized by stating that individual properties are known as, or presented by, properties of objects or values of properties.

value p of P), are reported to be the same individual property, $p_1 = p_2$ and therefore $P[a] = p$, as the result of the evaluation. Any Basic Evaluation Equation is then interpreted to be a mere identity from an *ontic* point of view, but a significant relation from an *epistemic* point of view.[15] This reveals the fundamental meaning of the Basic Evaluation Equation:

- the property $P[a]$ of the object a is an individual property p (the mass of any given object is a given mass);
- by means of a measurement the individual property p that was known as $P[a]$ is identified also as a given value p of P (the mass that was known as the mass of a given object is identified also as 1.2345 kg).

These two basic arguments need to be carefully presented, explained, and justified, and to this purpose the balance of this chapter and the next one are devoted.

5.2 Some clarifications about properties

The concept <property> has some ambiguities that we need to discuss before proceeding with our analysis.

5.2.1 Properties of objects as entities of the world

First, by considering properties of objects as associated with modes of interaction of the objects with their environments we acknowledge that properties of objects exist in the empirical world, and thus not only in our minds.[16] A given object generally has multiple modes of interaction with its environment, and each corresponds to a property of the object. In fact, against radical operationalism, it may be discovered that the same property is the cause of distinct modes of interaction.[17]

According to the tripartition introduced in Sect. 2.1, properties of objects are then, at least preliminarily, claimed *to be entities of the world*, not conceptual entities and not linguistic entities (as discussed further in Sect. 5.3). For example, that an object floats in water is a fact that can be observed or experimentally assessed, and is independent of the information that we may have on the object and its properties. That is, objects floated before Archimedes' explanation in terms of the relation between the weight and the shape of the object and the density of the water. In other words, a property of an object can be conceptualized and given a term, but it is not itself a concept or a linguistic expression. Of course, there can be disputable observations and mistaken reports of observations: what we assume here is just that at least some interactions are uncontroversially observed, and that there must be something in the empirical world, thus independent of our conceptions, which causes those interactions.

[15] This double meaning – an ontic identity that is a significant epistemic relation – is introduced in Box 5.1 and more extensively discussed in Sects. 5.3.2 and 6.4.

[16] The concept <property> is so general that this condition needs to be specified. For example, one could consider that the number 2 has the property of being even and of having 4 as its square: considering these as modes of interaction would be peculiar, at least because the very idea that numbers interact with something empirical is peculiar in turn. Hence, our analysis actually relates to *empirical* properties and *empirical* modes of interaction. We use the adjective "empirical" for referring to a feature of something in opposition to the possibility that that something is purely conceptual, informational, or linguistic (see also Sect. 2.2.1).

[17] According to Abraham Kaplan, "We do not first identify some magnitude, then go about devising some way to measure it. As operationists have long insisted, what is measured and how we measure it are determined jointly. Operationists may have given undue emphasis to the "how" as against the "what", but this emphasis is a healthy corrective to the naive idea that magnitudes can be conceived quite independently of procedures for determining their measure in particular cases." (1964: p. 177). Hence, we endorse such a "naive [!] idea", and consider that quantities (i.e., Kaplan's "magnitudes"), and properties more generally, *can be* identified by observing some effects and ascribing them to properties, thus still independently of procedures for measuring such properties.

There is an analogy in this between objects and properties (see Fig. 5.3). Concepts and linguistic terms can be associated with objects, such as rods and human beings, but rods and human beings remain something other than concepts and linguistic terms. Objects can exist without any associated concept or linguistic term (an obvious example being objects that existed before conscious beings evolved), and vice versa we can have concepts and linguistic terms of or about non-existing objects, as in the canonical cases of unicorns and phlogiston. Furthermore, while objects are subjected to empirical transformations (rods can rust, human beings grow old, etc.), concepts and linguistic terms of or about objects are unaffected by such transformations, but can be adjusted to better match objects (concepts and linguistic terms do not rust, but the concept of a rusted object is different from the concept of a polished object, etc.). Further, what can be properly defined is the concept of an object, not the object as such: objects are manipulated, designed, assembled, identified, etc., but not defined. By analogy, concepts can be provided of properties of objects, such as the length of a rod and the reading comprehension ability of an individual, but lengths of given rods and reading comprehension abilities of given individuals are not supposed to be themselves concepts. Indeed, while properties of objects can be subjected to empirical transformations (the length of a rod can change, the reading comprehension ability of an individual can improve, etc.), the concepts of properties of objects are unaffected by such transformations. Importantly, then, when expressions such as "unit definition" (e.g., throughout the SI Brochure, BIPM, 2019) and "measurand definition" are used (e.g., in the definition of <definitional uncertainty> given by the *International Vocabulary of Metrology* (VIM) (JCGM, 2012: 2.27), they are just shorthands for "definition of the concept of the unit" and "definition of the concept of the measurand", or, more operationally, "definition of the mode of identification of the unit" and "definition of the mode of identification of the measurand".

An exception to this is found in objects and properties of objects whose existence is dependent on (usually but not always shared) belief and social agreement, such as money, limited liability corporations, marriages, and beauty, which are often referred to as "social constructs" (see, e.g., Searle, 1995) to emphasize the role of human intentionality in their existence. But even in such cases there is a distinction between the object or property and the concepts one may have of it (for example, one may have a concept of money, separately from having money), and so the other comments given here about the distinction between objects and concepts remain applicable. In Sect. 6.6 we further discuss the existence of these kinds of properties.

A summary can be depicted as in Fig. 5.3, adapted from ISO (2009: 5.4.1, where the term "characteristic" is used to denote concepts of properties of objects).

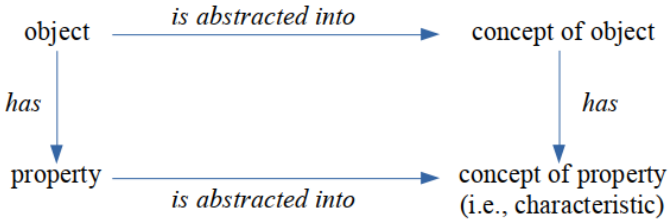


Fig. 5.3 Graphical representation of the relations between objects, properties, and their concepts

In this context, the condition that properties of objects are associated with modes of interaction is not obvious. In Sect. 3.4.1 we mentioned the "hage" of a person, defined as the product of her height and age, and presented as an exemplary case of a supposedly non-existing property of human beings (Ellis, 1968: p. 31). In other words, this is a definition of a perfectly legitimate concept, but it might not correspond to any empirical property. However, this is not something that can be taken for granted:

sooner or later, it might happen that a hage-related interaction of human beings is discovered. Again, the problem is meaningful because it relates to the claim that a property does exist as such, whereas that the concept of hage exists is only a matter of someone having ever thought about it, possibly even as a counterexample in a discourse about properties and their existence.

While the arrow in Fig. 5.3 points from property to concept of property, indicating that a concept of an empirical property must be derived from the property itself, the historical relationship may have well gone the other way; that is, one might develop a concept of a property before the conceived property is found empirically. This sort of historical sequence is far too many centuries in the past to be observed for common physical properties like length and weight, but this is not so for properties such as energy and temperature, and for many properties in the human sciences.

5.2.2 Properties and predicates

In the philosophical tradition, and in formal logic in particular, a property is what a predicate designates,[18] and is therefore a Boolean entity that either applies or does not apply to a given object, or, as more commonly said, that a given object either has or does not have. The distinction between predicates (as well as characteristics) and properties is effectively depicted as in Fig. 5.4.

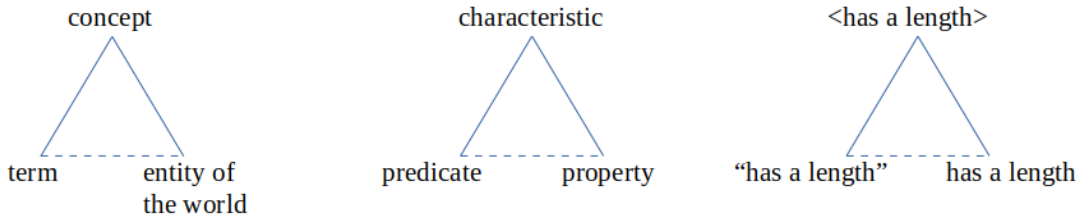


Fig. 5.4 The semiotic triangle (as in Fig. 2.1) applied to properties in the sense of formal logic

For example, properties in this sense are designated by the predicates “has a length”, “is longer than one metre”, and “is 1.2345 m long”: for any given object *a* that is the subject of these predicates, it is assumed that either it has a length or it has not, and so on. If a property in the sense of formal logic (hereafter designated $P^{\#}$ for maintaining a notational distinction with the properties as considered in measurement science, P) applies to an object, and therefore the corresponding predicate applied to a term that designates the object is true, we say that the object *has* that property: a rod *a* has the property of having a length, can have the property of being longer than one metre, etc. The proposition that the rod *a* is longer than one metre is then written as[19]

$$is\_longer\_than\_one\_metre(rod\ a) = \text{true}$$

whereas for example it might be that

$$is\_longer\_than\_one\_metre(screw\ a') = \text{false}$$

[18] There are many excellent books that can be used as reference on formal logic. The textbook by Hodges (1977), for example, is interesting for its explicit emphasis on the relations between natural languages and logic and the absence of required mathematical pre-competences.

[19] The relations $P^{\#}(a)$ = true and $P^{\#}(a)$ = false are usually written $P^{\#}(a)$ and not($P^{\#}(a)$) for short respectively. In what follows, we use the same symbols and expressions to denote properties and the corresponding predicates. This notational choice, of using the same symbol for a property and the mathematical entity that models the property, is usual, and for example the *Guide to the expression of uncertainty in measurement* (GUM) adopts it with this explicit justification: “For economy of notation, in this Guide the same symbol is used for the physical quantity (the measurand) and for the random variable that represents the possible outcome of an observation of that quantity.” (JCGM, 2008: 4.1.1, Note 1). Nevertheless, such a choice is sometimes a possible source of confusion, that should be avoided: even when properties are not notationally differentiated from their concepts and expressions, as previously noted, properties are not concepts and are not expressions.

There is an ambiguity about the very concept that an object does not have a property. For example, if we consider now the object *water a″ = the water in a given glass*, should we simply accept that

is_longer_than_one_metre(*water a″*) = false

despite its obvious difference with the previous case? Hence the problem arises of maintaining a distinction between the case of objects that do not have a property $P^{\#}$ but could have it and the case of objects that even in principle cannot have a property $P^{\#}$. This is based on the idea that only in the first case can $P^{\#}$ be experimentally assessed on the object, and thus that the second case would be better reported as

is_longer_than_one_metre(*water a″*) = undefined

We can account for this difference by restricting the application of each property $P^{\#}$ to a given set of objects, called the *domain* of $P^{\#}$. Hence the object *screw a′* belongs to the domain of the property *is_longer_than_one_metre*, and might not have the property, whereas the object *water a″* does not belong to the domain of the property, because trying to assess whether some amount of water in a glass is longer than one metre is meaningless. The distinction between physical properties and psychosocial properties is then usually and first a distinction of domain: a rod has no reading comprehension ability (i.e., the domain of reading comprehension ability does not include rods), and a company has no length (i.e., the domain of length does not include companies). In summary, for each property $P^{\#}$ the set of objects is split into three subsets: the subset of objects that actually have $P^{\#}$, the subset of objects that may have $P^{\#}$ but do not actually have it, and the subset of objects that cannot have $P^{\#}$. The identification of the domain of a property, i.e., the union of the first two subsets, can be considered an essential component of the knowledge of that property.[20]

5.2.3 Properties and relations

In the philosophical tradition, and again in formal logic in particular, a distinction is also maintained between properties and relations, where the former are features of (i.e., apply to) single objects and the latter are features of two or more objects, and all are designated by predicates with either one or two or more arguments respectively. For example, ordering is a relation between pairs of objects – if a is less than a' then *order*(a, a') = true (more usually written $a < a'$) – and betweenness is a relation between triples of objects – if a is greater than a’ and less than a’’ then *between*(a, a', a'') = true (more usually written $a' < a < a''$). In this sense any physical quantity that is relative to a reference system is a relation, as for example is the case of the speed of an object, which is not a property of the object but a relation between the object and the system in reference to which speed is considered. Hence, according to this terminology, while *has_a_given_length* is a property that an object can have, *has_a_given_speed* is a relation, e.g., *has_a_given_speed*(*rod a*, *reference system b*).[21]

This distinction is not usually maintained in measurement science, in which the terms “property” and “quantity” are usually applied both to what would be considered properties and relations in formal

[20] If the position is assumed that *every* predicate designates a property (in the sense of formal logic), things can become tricky. Consider, e.g., the predicate “is a length”: if *is_a_length*(*a*) = true, then it is because *a* is a property, and it is in fact a length. The domain of *is_a_length* is then a set of properties – so that *is_a_length*(*a given rod*) is undefined, not false – and *is_a_length* is a higher-order property, i.e., a property of properties. It is doubtful that such kinds of properties can be assessed via empirical interactions (on the other hand, *is_a_quantity* is an example of a second-order property, which instead admits an empirical validation – see the related discussion in Sect. 6.3.2).

[21] This is what the Galilean relativity principle asserts. According to Einstein’s relativity theory, the length of a body observed from a frame of reference in motion with respect to the body depends on the relative velocity of the two systems, so that in this view length is also a relation. Moreover, even in classical physics *is*_1.2345_m_*long*(*a*) could be re-interpreted as the relation *is*_1.2345-*fold_long*(*a*, *s*) between the object *a* and any measurement standard *s* which materializes the metre: from this perspective, *all* ratio properties treated in measurement can be interpreted as relations.

logic. For a property to change it is then sufficient that one object, to which the property applies, changes.[22]

We accept this terminological custom here, and – consistently with the current edition of the VIM (JCGM, 2012: 1.1) – use the term "property" for relations as well.[23] With this convention, the difference between properties in the sense of formal logic and properties in the sense of measurement science can be analyzed.

5.2.4 From properties of formal logic to properties of measurement science

Since a Basic Evaluation Equation such as

$$\mathit{length}[\mathit{rod}\ a] = 1.2345\ \text{m}$$

can be rewritten in the predicative form as

$$\mathit{is}\_1.2345\_\text{m}\_\mathit{long}(\mathit{rod}\ a) = \text{true}$$

one could conclude that these expressions convey exactly the same information. This is not the case, and a consideration of the differences allows us to highlight some fundamental features of properties (in the sense of measurement science, the meaning to which we implicitly refer henceforth).

Consider the three (logical) equations:

$$\mathit{is}\_1.2345\_\text{m}\_\mathit{long}(\mathit{rod}\ a) = \text{true} \quad (1)$$

$$\mathit{is}\_2.3456\_\text{m}\_\mathit{long}(\mathit{rod}\ a) = \text{true} \quad (2)$$

$$\mathit{is}\_3.4567\_\text{kg}\_\mathit{heavy}(\mathit{rod}\ a) = \text{true} \quad (3)$$

While eq. (1) and eq. (3) can hold at the same time, eq. (1) and eq. (2) cannot. However, the predicative form $P^{\#}(\mathit{object})$ is unable to prevent both eq. (1) and eq. (2) from being asserted as true at the same time. Indeed, consider rewriting the three predicates as $P^{\#}_1$, $P^{\#}_2$, and $P^{\#}_3$ respectively: how could one know that, for a given x, both $P^{\#}_1(x)$ and $P^{\#}_3(x)$ can be true but that if $P^{\#}_1(x)$ is true then $P^{\#}_2(x)$ must be false?

In order to acknowledge that eq. (1) and eq. (2) are incompatible, the involved properties must be recognized as having some sort of internal structure, such that the equations could be rewritten in a parametric form as

$$\mathit{is\_long}_{1.2345\_\text{m}}(\mathit{rod}\ a) = \text{true} \quad (1_a)$$

$$\mathit{is\_long}_{2.3456\_\text{m}}(\mathit{rod}\ a) = \text{true} \quad (2_a)$$

$$\mathit{is\_heavy}_{3.4567\_\text{kg}}(\mathit{rod}\ a) = \text{true} \quad (3_a)$$

or in the relational form

$$\mathit{is\_long}(\mathit{rod}\ a, 1.2345\ \text{m}) = \text{true} \quad (1_b)$$

[22] This avoids the need of specifically dealing with the so called *Cambridge changes*, "such as when I change from having "non-brother" true of me to having "brother" true of me, just when my mother gives birth to a second son", the problem being of course that "it might seem faintly paradoxical that there need be no (other) changes in me (height, weight, colouring, memories, character, thoughts) in this circumstance" (Mortensen, 2020: chapter 2). Consider an example closer to our context, like

$\mathit{is\_at\_a\_distance\_of}\_1.2345\_\text{m}\_\mathit{from}(\text{body}\ a, \text{reference}\ b) = \text{true}$

If instead we assumed that the property is

$\mathit{is\_at\_a\_distance\_of}\_1.2345\_\text{m}\_\mathit{from\_reference\_b}(\text{body}\ a) = \text{true}$

then changes of the position of b would need to be considered also as changes of a property of a, even though a itself did not move.

[23] Sometimes the term "attribute" is used to encompass properties and relations. This was plausibly the choice of the second edition of the VIM, which defines <(measurable) quantity> as an "attribute of a phenomenon, body or substance that may be distinguished qualitatively and determined quantitatively" (BIPM et al., 1993: 1.1).

is_long(*rod a*, 2.3456 m) = true (2_b)

is_heavy(*rod a*, 3.4567 kg) = true (3_b)

that under the functional condition of uniqueness – for all x, $P^{\#}(x, y_1)$ = true and $P^{\#}(x, y_2)$ = true implies $y_1 = y_2$ – corresponds to the more usual functional form

long[*rod a*] = 1.2345 m (1_c)

long[*rod a*] = 2.3456 m (2_c)

heavy[*rod a*] = 3.4567 kg (3_c)

where "long" and "heavy" are not predicates anymore (as used in this way "long" is then different from the predicate "is long"[24]), but examples of what Rudolf Carnap called *functors* (1937: p. 14; a discussion on predicates and functors in the context of measurement is in Mari, 1996). In short, once the set of possible values of the parameter x is given, one functor, "long", which maps objects to values, corresponds to the whole set of predicates "is $\text{long}_{x\ \text{m}}$". Hence, just as predicates are the linguistic counterparts of properties and relations in the sense of formal logic, functors are the linguistic counterparts of properties in the sense of measurement science.[25]

With a formalization based on functors, the incompatibility of eq. (1_b) and eq. (2_b) becomes explicit. However, this cannot be justified on the basis of the linguistic fact that the functor "long" is the same in these equations: they remain incompatible even if in eq. (2_b) "long" is translated into another language, e.g., into the Italian "lungo". Such an incompatibility is an empirical fact, which calls for a justification, to be developed in the sections that follow. Interestingly, the basics of a measurement-oriented ontology and epistemology of properties can be first developed without recourse to values of properties, which will deserve a specific analysis on their own.

5.2.5 Context dependence of properties

Our analysis of properties and objects is grounded on the basic assumption that objects can persist in space and time even if one or more of their properties change.[26] In particular, by indexing properties

[24] Admittedly, a form such as "*long*[*rod a*]" is clearly awkward, and is introduced here only as an intermediate step from properties in the sense of logic, e.g., *is long*[*rod a*], to properties in the sense of measurement science, e.g., *length*(*rod a*).

[25] There is in fact another functional form for conveying the information brought by a Basic Evaluation Equation:

*long_in*_m[*rod a*] = 1.2345

*long_in*_m[*rod a*] = 2.3456

*heavy_in*_kg[*rod a*] = 3.4567

We further discuss it in particular in Sect. 6.2.2, in the context of the analysis of the way representational theories of measurement deal with values of properties.

[26] For example, on the one hand, at least in the broadly Western tradition, each of us admits our own persistence in time as individuals even though we change, say, our height and weight and, less trivially, our cognitive abilities. On the other hand, an object *can* change to another one if one or more of its properties change, as in the case of an informous amount of clay that is modeled and finally becomes a jar. An extreme case of the dilemma of the conditions of object persistence is known since the classical world as the *Theseus paradox* (see Korman, 2016: 2.4): does a ship remain the same even if, one by one, all its wooden boards are substituted? And therefore: is an object characterized by the matter of which it is constituted, or by its shape, or by which combination of the two? The assumption of some basic persistence is intrinsic to our concept of object. Even just imagining how to avoid it is challenging. In one of his tales, "Funes el memorioso" ("Funes the Memorious"), Jorge Luis Borges (1944) tried imagining what it would be like to avoid it, by telling of an individual of prodigious memory: "In the seventeenth century, Locke postulated (and condemned) an impossible language in which each individual thing – every stone, every bird, every branch – would have its own name; Funes once contemplated a similar language, but discarded the idea as too general, too ambiguous. The truth was, Funes remembered not only every leaf of every tree in every patch of forest, but every time he had perceived or imagined that leaf.". Compare this with: "All is impermanent, because all is in a state of perpetual change. A thing does not remain the same during two consecutive ksanas (the ksana being the shortest period of time in Buddhism). It is because things transform themselves ceaselessly that they cannot maintain their identity, even during two consecutive ksanas." (Nhat Hanh, 1974: p. 35).

of objects by time instant (so that, for example, *long*[*a*, *t*] is the property designated by the functor "long" that the object *a* has at time instant *t*), a property-related comparability criterion is given such that, for distinguishable time instants *t* and *t′*, it may happen that

$$long[a, t] \approx long[a, t']$$

$$heavy[a, t] \not\approx heavy[a, t']$$

where ≈ denotes indistinguishability with respect to the given criterion:[27,28] it is a situation in which the object *a* has maintained its individuality from *t* to *t′* because in particular its being long has not changed,[29] while, plausibly among other properties, its being heavy has changed.

A complementary basic assumption is that property-related comparability is applicable not only to the same object in different time instants, but also to different objects, in the same or different time instants: *objects are comparable via the comparison of their properties*. Hence, it may happen, for example, that

$long[a, t] \approx long[b, t]$ (*a* and *b* are synchronously indistinguishable in their being long)

$long[a, t] \approx long[b, t']$ (*a* and *b* are asynchronously indistinguishable in their being long)

but also, of course, that

$long[a, t] \not\approx long[b, t]$ (*a* and *b* are synchronously distinguishable in their being long)

$long[a, t] \not\approx long[b, t']$ (*a* and *b* are asynchronously distinguishable in their being long)

More generally, properties are acknowledged to be mutually related, so that the property of being long of an object may change not only with time but also, for example, with the temperature of the object, the pressure of the surrounding environment, etc. Hence, there is a context *c* that influences (and is influenced by) *long*[*a*]: we designate by *long*[*a*, *c*] the property of being long of the object *a* in the context *c*, with the understanding that time is part of the context, or possibly *long*[*a*, *c*, *t*] in order to emphasize the time instant when the property is taken into account.

The relation of experimental indistinguishability of properties of objects deserves some more analysis.

5.2.6 Indistinguishability of properties of objects

Let us assume that the indistinguishability of two properties of objects has been observed, as in the case

$$long[a] \approx long[b]$$

[27] By acknowledging that time instant is also a property (of the reference system shared by the considered objects), the condition that *t* and *t'* are distinguishable time instants should be written $t \not\approx t'$, not $t \neq t'$, thus emphasizing that two time instants could be indistinguishable (because, e.g., they are within 1 nanosecond of one another and the quality of the available instrumentation is not able to detect this difference) even though they are not necessarily exactly the same.

[28] A relation such that object *a* at time instant *t* and object *a*' at time instant *t*' are indistinguishable in their being long may be differently interpreted. If objects are considered to be entities with time instances, then the relation could be written $long[a(t)] \approx long[a'(t')]$ (see Mortensen, 2020: chapter 5). By focusing even more explicitly on objects and considering *to be long* as a feature of the way objects are compared, the relation could be written instead $a(t) \approx_{long} a'(t')$. The focus in measurement science on properties – what is measured is the property of an object, not an object *per se* – justifies our choice of adopting the form $long[a] \approx long[a']$, and possibly its time explicit version $long[a, t] \approx long[a', t']$ whenever appropriate. The alternative position of focusing on objects is taken in particular in the representational theories of measurement, which usually develop their formalization on empirical relations among objects; see, e.g., how the concept <relational structure> is introduced in Krantz et al. (1971: p. 8).

[29] "Its being long has not changed" is awkward phrasing, and could be changed to the more usual "its length has not changed". We maintain it at the moment, given that the relation of the adjective "long" with the corresponding noun "length" is discussed afterwards.

How should such a relation be interpreted?[30] An aspect of the issue pertains to the fact that property comparisons relevant to measurement are empirical activities, as discussed in Sect. 2.3, and that any such activity is usually affected by factors that prevent its ideal realization. Hence, even when two properties are found to be indistinguishable, a more specific comparison might reveal that they are *not equal*, but only *similar*.

Like any generic similarity, indistinguishability is

- reflexive (any property is indistinguishable from itself: $P[a] \approx P[a]$) and
- symmetric (if two properties are indistinguishable, then the order in which they are considered is immaterial: $P[a_i] \approx P[a_j]$ if and only if $P[a_j] \approx P[a_i]$), but
- not transitive (given three properties, from the facts that the first and the second are indistinguishable and that the second and the third are indistinguishable, the conclusion that also the first and the third are indistinguishable does not follow: $P[a_i] \approx P[a_j]$ and $P[a_j] \approx P[a_k]$ do not imply that $P[a_i] \approx P[a_k]$).

The non-transitivity of indistinguishability has at least one critical consequence: properties of objects could not be consistently represented, or even named, in any sufficiently simple form. Indeed, the observation that $P[a_i]$ and $P[a_j]$ are indistinguishable would lead one to represent them with the same symbol, and the observation that $P[a_j]$ and $P[a_k]$ are also indistinguishable would lead one to represent $P[a_k]$ with the same symbol as $P[a_j]$, and therefore as $P[a_i]$; but since $P[a_i]$ and $P[a_k]$ could be instead distinguishable, this would lead to the situation in which distinguishable properties are represented by the same symbol, a case of homonymy and therefore information loss in representation. Furthermore, since the comparison can be iterated, there might be a sequence of objects $a_1, a_2, ..., a_n$ such that $P[a_1] \approx P[a_2]$ and $P[a_2] \approx P[a_3]$ and ... $P[a_{n-1}] \approx P[a_n]$ even though, for each i, $P[a_i] \not\approx P[a_{i+2}]$, with the consequence that all comparable but mostly distinguishable properties of objects are represented by the same symbol, and therefore that the representation conveys no information at all.[31]

This issue does not seem to have a general solution better than provisionally assuming the transitivity of indistinguishability, and therefore modeling the comparison of properties of objects as an equivalence relation, which could be then discovered to be not such by means of further and more refined comparisons. As we discuss below, providing information on properties of objects in terms of traceable values is the specific solution to this problem adopted in measurement science (see also Mari & Sartori, 2007).

5.3 A philosophical interlude

The analysis developed so far about properties has been mainly technical, aimed at setting a conceptual framework for interpreting the Basic Evaluation Equation

property of a given object = value of a property

In fact, such an interpretation has an unavoidable, though possibly only implicit, philosophical background. Given a Basic Evaluation Equation such as

[30] In what follows, it is immaterial whether the comparison is synchronous or asynchronous, and whether it depends on the context: therefore the reference to time and context is usually omitted.

[31] This is the paradox known as *sorites*, a term which derives from the Greek word σωρός, "soros", meaning <heap> (see Hyde & Raffman, 2018). The classical way to present it is in terms of logical properties, for example as follows. Let a_n be a set of n grains of wheat. Of course a_0 is not a heap, i.e., $is\_heap(a_0)$ = false. Moreover, a_n and a_{n+1} are indistinguishable in their being heaps, in the sense that if a set is not a heap, adding a grain to it does not make it a heap, i.e., if $is\_heap(a_n)$ = false then $is\_heap(a_{n+1})$ = false. Then starting from the first clause and by the repeated application of the second clause the conclusion is reached that $is\_heap(a_n)$ = false no matter how large is n.

length[*rod a*] = 1.2345 m

one may claim, for example, that this is just a sophisticated linguistic shorthand for reporting something about the relation between two objects, *rod a* and an object that in some primitive sense is attributed "to be one metre", or possibly "to have one metre", but that there is nothing in the world that corresponds to <being one metre> or <having one metre>, let alone the <length of *rod a*> and <one metre>. Accordingly, properties of objects and values of properties would be nothing more than conceptual tools created to organize our knowledge of the world.[32] However, a Basic Evaluation Equation could be instead interpreted literally, as reporting a relation between entities – properties of objects and values of properties – that exist in the world, of course with respect to a given mode of existence, as discussed in Box 5.1.

The subject is so fundamental for any ontology (see references in Footnote 5) that we will not feign to provide any original contribution to this discussion: this chapter is concluded simply by discussing why we support a realist position about individual properties, while postponing to Sect. 6.6 the more complex analysis of the existence of general properties.

5.3.1 Do individual properties exist?

Even if the experimental issues regarding the indistinguishability of properties of objects are somehow settled, there remains a general ontological problem. Let us suppose that the repeated application of the best available means of comparison has been unable to find any difference between the comparable properties of two objects, so that according to the available information the indistinguishability relation $P[a_i] \approx P[a_j]$ could be in fact considered an equality, $P[a_i] = P[a_j]$: how should such a relation be interpreted?

There are two, basically alternative, answers to this question (see Orilia & Swoyer, 2020).[33]

- According to one position, no matter how similar $P[a_i]$ and $P[a_j]$ are, they are still distinct entities, as the objects a_i and a_j are distinct, because any property of an object is by definition a property that only that object has and can have. In this sense, the equality $P[a_i] = P[a_j]$ is simply a convenient notation for an equivalence $P[a_i] \cong P[a_j]$: while the objects a_i and a_j have indistinguishable modes of interaction with the environment, their properties remain different because they are *of* different objects. A plausibly unavoidable consequence of this position is that an object a at the time instant t_i and the same object at any other time instant t_j – as introduced in Sect. 5.2.5 – cannot have the same property in turn, i.e., $P[a, t_i]$ and $P[a, t_j]$ *must* be distinct entities.

  Under this assumption, the ontology of properties is then straightforward, given that it may include only properties of objects at time instants as fundamental entities,[34] and it assumes their relation of indistinguishability as primitive. Thus, there are no individual properties independent of objects at time instants, and what we consider lengths, given reading comprehension abilities, and so forth are only concepts that we create to help organize our experience of and communication about the world. The price to be paid is that each property

[32] Depending on how radical (or coherent?) this position is, one could say the same about objects: the world *is* in principle an undifferentiated blob, and objects are only conceptual constructions we provide for understanding the world.

[33] Note that this alternative is presented with no reference to values of properties, which are instead discussed in Chap. 6.

[34] With or possibly even without objects: in an extreme position, objects could be just considered as the bundles of their properties – see Maurin (2018) (and since properties change as time passes, this also means that objects have no persistence: I am not the same person as I was one second ago, but we are only equivalent, according to a more or less complex criterion of equivalence; this might be an interpretation of the radical impermanence mentioned in Footnote 26 about Buddhism).

of each object at each time instant is a distinct entity, and therefore that at each new instant a bunch of properties is created, so that an ever increasingly growing multitude of properties exists. This is an instance of a position that in the philosophical tradition is often called *nominalism*.

- According to another position, properties of objects are such that distinct objects can have one and the same property. Hence the indistinguishability of $P[a_i]$ and $P[a_j]$ suggests that they could be indeed the same property. In this sense, $P[a_i] = P[a_j]$ is the theoretical counterpart of $P[a_i] \approx P[a_j]$: the fact that the objects a_i and a_j have indistinguishable modes of interaction with the environment supports the hypothesis that they have the same property. In the lexicon of ontology, individual properties are then *universals*, that can be instantiated in, i.e., exemplified by, one or more objects.

  The ontology of properties in this case might be considered then more complex than in the previous case, given that it includes universal entities, such as any given length and any given reading comprehension ability, together with particulars such as the length of any given rod and the reading comprehension ability of any given individual. However, this position avoids assuming that each property of each object in each instant is a distinct entity, and accounts for the relation of indistinguishability of properties of objects in a simple way: two properties of objects are indistinguishable if they identify the same universal property, or if they identify distinct universal properties which prove to be indistinguishable according to the available empirical means. This is an instance of a position that in the philosophical tradition is often called *realism*.

The two positions share the common condition that properties of objects are particular entities, spatiotemporally situated. Whether things such as lengths and reading comprehension abilities are concepts, as nominalists assert, or universals that exist in an abstract world independently of the knowledge that we have of them, as realists assert, is an ontological alternative that affects the interpretation of some key components of measurement, in particular the meaning of the Basic Evaluation Equation, but cannot be decided by experimental activity.[35] Each position can be translated into the other one safely (and usually unproblematically, though sometimes in a cumbersome way – see the examples in Footnote 35; see also Mari & Giordani, 2012: p. 762).

We believe that most scientists, technologists, and practitioners, in both the physical and human sciences, are, perhaps unconsciously, at least moderate realists about properties, in that they consider individual properties such as lengths and reading comprehension abilities to exist in some sense (see also the discussion in Sect. 6.6). If explicitly asked, they might plausibly accept the more complex ontology in which properties are considered to be universals, given that the adoption of the more

[35] The alternative between realism and nominalism relates not only to properties, as realism and nominalism clash about the existence of universals as such. Take these two examples: "the steam engine was invented at the end of the seventeenth century" and "the tiger is an endangered species". For a nominalist they cannot be literally true, because "the steam engine" and "the tiger" do not refer to anything in the world, but only to concepts we adopt to organize our knowledge. She would explain the meaning of the two sentences by considering them to be shorthands for something like "the first object that we presently conceptualize as <steam engine> was invented at the end of seventeenth century" and "the current number of objects that we conceptualize as <tiger> is less than a given threshold", or even more explicitly "the extension of <steam engine> was empty before the end of seventeenth century" and "the cardinality of the current extension of <tiger> is less than a given threshold". This reduction strategy may become cumbersome. For example, in the case of "Shakespeare's works include 39 plays", the nominalist would claim that the quantification is on concepts, so that, e.g., *A Midsummer Night's Dream* does not exist as such, but it is only a concept by means of which we identify a subset of the objects (paper volumes, digital files, theater performances, etc.) with the property of having Shakespeare as their author. Assessing the truth of the sentence would then require first assessing at least implicitly an equivalence criterion between such disparate entities, and then counting the number of the so obtained equivalence classes.

complex ontology does not affect day to day practice, and simplifies the reporting of experimental results by at least provisionally accounting for the observed indistinguishability of properties of objects in terms of their equality, such that properties of objects can be included in relations interpreted as actual equations. Furthermore, this ontology allows for a more flexible treatment of properties, in particular by admitting the possibility of formally handling properties that might currently not be instantiated by any existing object (and maybe even properties that are known not to be instantiated by any object, such as lengths greater than the diameter of the universe and masses greater than the mass of the universe).[36] Finally, while a realist ontology has a greater categorical complexity, it spares the nominalist requirement of an immensely great number of properties, immensely growing at each time instant with the creation of new properties. For these reasons, we maintain here the position that individual properties are universals.

5.3.2 Individual properties as universals: an explanation

The idea that individual properties are universals is conceptually sophisticated: how can it be that indistinguishable properties of *distinct* objects may correspond in fact to the *same* individual property? Let us consider a mathematical relation such as $\sum_i 1/(i\ 2^i) = \ln(2)$, where i is an integer ranging from 1 to infinity, an equation which is known to be true, given that both $\sum 1/(i\ 2^i) = 0.693147...$ and $\ln(2) = 0.693147\ldots$. In terms of the involved numbers the relation $\sum_i 1/(i\ 2^i) = \ln(2)$ is not different from $0.693147... = 0.693147...$: but while the latter is a logical identity, which does not convey any information, the former implies some mathematical knowledge, so that in some respect the two entities, $\sum_i 1/(i\ 2^i)$ and $\ln(2)$, must be different. However, there is also a respect in which the equality actually holds, so that we can say that $\sum_i 1/(i\ 2^i) = \ln(2)$ is true, while, for example, $\sum_i 1/(i\ 2^i) = 2$ is false. This double interpretation – different but equal at the same time – accounts for the principled difference between $\sum_i 1/(i\ 2^i) = \ln(2)$ and $0.693147... = 0.693147...$, which can be explained in terms of the distinction between the *sense* and the *reference* of an expression (as noted in Box 5.1 above), where the sense of an expression is the concept it designates and the reference of an expression is the entity it refers to.[37]

Hence "$\sum_i 1/(i\ 2^i)$" and "$\ln(2)$"

- are different *expressions* (the former starts with a symbol of summation, the latter with the name of a function, and so on),
- with different *senses*, since they designate different concepts (the former is a series, the latter is a function evaluated in a given argument),
- but with the same *referent*, since they refer to the same mathematical object (the number $0.693147...$).

This is a possible interpretation of the relation $P[a_i] \approx P[a_j]$, and indeed the one we adopt here: when claiming, e.g., that the length of a_i is indistinguishable from (or even the same as) the length of a_j, we interpret this as the hypothesis that there is one individual length that is identified as the length of the two objects, and therefore is known in two different ways. In fact

[36] There is one more reason supporting realism about properties, related to the status of values of properties, a subject that we explore in Chap. 6. Just as a mention here, though obtained through the conventional definition of a unit, an entity such as 1.2345 m seems to have an existence independent of the knowledge that we have of it. In other words, values are not concepts.

[37] An analogous distinction is put between the intension and the extension of a concept (see Sect. 2.1). Hence, the intensions of the concepts $<\sum_i 1/(i\ 2^i)>$ and $<\ln(2)>$ are different, while their extension is the same, i.e., the number $0.693147...$ Note that intensions and extensions are sometimes attributed to terms too; see, e.g., Chalmers (2002).

- while the expressions "$P[a_i]$" and "$P[a_j]$" have different senses (because they convey information on properties of different objects),
- their referent could be the same, and actually is the same if the equation $P[a_i] = P[a_j]$ is true, i.e., they refer to the same individual length.

In other words, if $P[a_i] = P[a_j]$ is true, it is because there is one individual length which is a universal entity that both a_i and a_j have and therefore both $P[a_i]$ and $P[a_j]$ instantiate.

5.3.3 Do we really need properties?

The position developed in the previous section has some analogies with Bertrand Russell's conception of natural numbers: "Under what circumstances do two classes have the same number? The answer is, that they have the same number when their terms can be correlated one to one, so that any one term of either corresponds to one and only one term of the other. [...] When the relation holds between two [classes], those two [classes] have a certain common property, and vice versa. This common property we call their number. This is the definition of numbers by abstraction." (1903: p. 113-116). Accordingly, the natural number *n* is what all classes of *n* elements have in common, as identified by a one-to-one correspondence among their elements. Such a position is compatible with both

- an *extensionalist* position: the number *n* is *the class* of all classes of *n* elements, and
- an *intensionalist* position: the number *n* is *the property* that all classes of *n* elements share.

Hence, extensionalism about properties considers them to be nothing but "classes of the entities whose properties they are [, so that] for example, human *baldness* (or *being bald*) is to be identified with the class of all bald humans, while over the domain comprising all chunks of minerals, the property *crystalline* is the class of all crystalline rocks" (Rozeboom, 1966: p. 172). As Hilary Putnam discusses (1969), extensionalism on properties assumes that if, for all x, x is $P^{\#}$ if and only if x is $Q^{\#}$, then $P^{\#}$ and $Q^{\#}$ are the same property, in analogy to the fact that if the sets A and B have the same elements then they are one and the same set. This applies not only to the properties in the sense of formal logic, but also to the properties in the sense of measurement science. Any reference to a property would be then just a convenient shorthand for a given (although usually unknown) set: "the object a has a given length" would precisely and only mean "the object a belongs to a given set" – that is, the set of objects having the same length as a – and so on.[38] According to Joel Michell, the consequence is that "the ontology of modern science [comprises] material objects (or, alternatively, space-time points), sets of material objects, sets of sets of material objects, ..., *but no properties*" (1990: p. 305).[39]

Were the extensionalist objection against properties to be accepted, any discourse about the ontology and epistemology of properties should be deemed to be extrinsic, a purely linguistic

[38] This is about individual properties. There is also an extensionalist interpretation of general properties. For example, according to Earl Babbie (who uses a peculiar lexicon: "variable" for general property and "attribute" for value), "variables ... are logical sets of attributes. Thus, for example, male and female are attributes, and sex or gender is the variable composed of those two attributes. The variable occupation is composed of attributes such as farmer, professor, and truck driver. Social class is a variable composed of a set of attributes such as upper class, middle class, and lower class." (2013: p. 13). Along the same line, Michell claims that "the variable of length is simply the class of all lengths" (1990: p. 51).

[39] Indeed, representational theories of measurement usually formalize measurement as a mapping from objects to numbers, by thus not including properties in their framework: "the first problem ... the analysis of any procedure of measurement must consider ... is justification of the assignment of numbers to objects or phenomena" (Suppes & Zinnes, 1962: p. 3). See also Sect. 6.2.2. Interestingly, extensionalism models logical properties and properties of measurement science (i.e., what above we designated $P^{\#}$ and P respectively) in exactly the same way, as mappings from sets of objects to sets of values (whatever they are), the only difference being that the cardinality of the codomain of logical properties is 2 (Lawvere & Rosebrugh, 2003: 1.2).

shorthand reducible to a set-theoretical analysis. Again from William Rozeboom we take a general reply (1966: p. 172):

> What is objectionable about this [...] is that properties are really distinguishing features of the entities which possess them, as that in principle properties can be coextensive even though non-identical. Thus if all crystalline rocks were translucent and conversely, we should deny that crystallinity and translucency are the same property of rocks even though the class of crystalline rocks would be identical with the class of translucent rocks.

A more general and less hypothetical example is provided by any physical law which connects two quantities via a constant, as is the case of the Planck-Einstein relation $E = hv$, stating that the photon energy E is proportional to its frequency v, via the Planck constant h. According to this law, in the case of photons the individual property *having energy* e, for any given e, and the individual property *having frequency* e/h are coextensive (i.e., the set of photons a_i such that $E[a_i]$ = e and the set of photons a_j such that $v[a_j]$ = e/h are the same). Nevertheless, the quantities photon energy and photon frequency remain distinct. Moreover, though not logically contradictory, the idea that laws of physics establish relations among sets, and that products and powers of sets can be somehow considered (as in the case of kinetic energy, which depends on the square of velocity), is counterintuitive, to say the least. Finally, another well-known counterexample to the extensionalist position was provided by W. O. Quine (1951): according to the current knowledge, "creatures with a heart" and "creatures with a kidney" have the same extension, but their meanings are clearly not interchangeable.

This whole discussion seems to provide sufficient reasons to refuse the extensionalist objection, and more generally not to endorse a nominalist position, and to continue exploring a measurement-oriented ontology and epistemology of properties under a realist perspective, according to which the Basic Evaluation Equation conveys information about the relation between entities which have their own modes of existence in the world. On this basis, in the following chapter we develop the position that individual properties are universals by framing it in a metrological framework, in which values of properties also play a significant role, and then broaden the picture in order to consider the very problem of the existence of general properties.

References

Babbie E. (2013). *The practice of social research* (13th ed.). Belmont: Wadsworth.

Baron, S., Copley-Coltheart, R., Majeed, R., & Miller, K. (2013). What is a negative property?. *Philosophy, 88(1)*, 33–54.

Borges, J. L. (1944). *Ficciones* (in Spanish). First English translation, 1954.

Bunge, M. (1977). *Treatise on Basic Philosophy.* (Vol. 3. Ontology I: The furniture of the world). Dordrecht: Reidel.

Carnap, R. (1937). *The logical syntax of language*. London: Kegan. Reprinted by Routledge, London, 2000.

Chalmers, D. J. (2002). On sense and intension. In J. Tomberlin (Ed.), *Philosophical Perspectives 16: Language and Mind* (pp. 135–182). Oxford: Blackwell.

de Boer, J. (1995). On the history of quantity calculus and the International System. *Metrologia, 31*, 405–429.

Dennett, D. (2017). *From bacteria to Bach and back: The evolution of minds*. New York: Norton.

Dybkaer, R. (2004). An ontology on property for physical, chemical, and biological systems. *APMIS, 112* (Suppl. 117), 1–210. Retrieved from ontology.iupac.org

Ellis, B. (1968). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

Emerson, W. H. (2008). On quantity calculus and units of measurement. *Metrologia, 45*, 134–138.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik, 100*, 25–50 (English translation by M. Black, On Sense and Reference, in the public domain: en.wikisource.org/wiki/On_Sense_and_Reference).

Heil, J. (2003). *From an ontological point of view*. Oxford: Clarendon Press.

Hodges, W. (1977). *Logic*. New York: Penguin.

Hyde, D., & Raffman, D. (2018). Sorites paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/sorites-paradox

International Bureau of Weights and Measures (2019). *The International System of Units (SI), 9th ed.* Sèvres: BIPM (www.bipm.org/en/si/si_brochure).

International Bureau of Weights and Measures (BIPM) and other six International Organizations (1993). *International Vocabulary of Basic and General Terms in Metrology (VIM)* (2nd ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML), International Union of Pure and Applied Chemistry (IUPAC), the International Union of Pure and Applied Physics (IUPAP).

International Organization for Standardization (2022). *ISO 704:2022, Terminology work – Principles and methods* (4th ed.). Geneva: ISO.

International Organization for Standardization (2022). *ISO 80000-1:2022, Quantities and units – Part 1: General* (2nd ed.). Geneva: ISO.

Joint Committee for Guides in Metrology (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM (www.bipm.org/en/publications/guides/gum.html).

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)*. Sèvres: JCGM, 3rd ed (2008 version with minor corrections) (www.bipm.org/en/publications/guides/vim.html).

Korman, D. Z. (2016). Ordinary objects. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/ordinary-objects

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (vol. 1). New York: Academic Press.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52(2)*, 161–193.

Lawvere, F. W., & Rosebrugh, R. (2003). *Sets for mathematics*. Cambridge: Cambridge University Press.

Loux, M. J., & Crisp, T. M. (2017). *Metaphysics – A contemporary introduction*. New York: Routledge.

Mari, L. (1996). The meaning of "quantity" in measurement. *Measurement, 17*, 127–138.

Mari, L. (1997). The role of determination and assignment in measurement. *Measurement, 21,* 79–90.

Mari, L. (2017). Toward a harmonized treatment of nominal properties in metrology. *Metrologia, 54*, 784–795.

Mari, L., & Giordani, A. (2012). Quantity and quantity value. *Metrologia, 49*, 756–764.

Mari, L., & Sartori, S. (2007). A Relational Theory of Measurement: traceability as a solution to the non-transitivity of measurement results. *Measurement, 40*, 233–242.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770.

Maurin, A. S. (2018). Tropes. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/tropes

Maxwell, J. C. (1873). *A treatise on electricity and magnetism*. Oxford: Oxford University Press.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale: Erlbaum.

Mortensen, C. (2020). Change and inconsistency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/change

Nhat Hanh, T. (1974). *Zen keys*. New York: Anchor Books.

Orilia, F., & Swoyer, C. (2020). Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/properties

Possolo, A. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST measurement results*. NIST Technical Note 1900. Retrieved from www.nist.gov/publications/simple-guide-evaluating-and-expressing-uncertainty-nist-measurement-results

Price, G. (2001). On the communication of measurement results. *Measurement, 29*, 293–305.

Putnam, H. (1969). On properties. In N. Rescher (Ed.). *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday* (pp. 235–254). New York: Springer. Reprinted in *Mathematics, matter and method: Philosophical Papers* (Vol. I). Cambridge: Cambridge University Press, 1975.

Quine, W. V. O. (1948). On what there is. *Review of Metaphysics, 2(5)*, 21–38. Reprinted in From a logical point of view. Harvard: Harvard University Press, 1953.

Quine, W. V .O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*, 20–43.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16*, 170–233.

Russell, B. (1903). *The principles of mathematics*. London: Bradford & Dickens.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic, 23(2)*, 113–128.

Searle, J. (1995). *The construction of social reality*. Oxford: Oxford University Press.

Suppes, P., & Zinnes, J. L. (1962). *Basic measurement theory*. Technical Report n. 45, Institute for Mathematical Studies in Social Sciences. Stanford, CA: Stanford University.

Varzi, A. (2019). Mereology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/mereology

Weatherson, B., & Marshall, D. (2018). Intrinsic vs. extrinsic properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/intrinsic-extrinsic

Wilson, J. (2017). Determinables and determinates. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/determinate-determinables

Chapter 6.
Values, scales, and the existence of properties

This chapter aims to expand on the ontological and epistemological analysis of properties introduced in the previous chapter, with a discussion of three fundamental issues for measurement science. Restarting from the distinction between general and individual properties, the first is about the nature of *values* of quantities and more generally of properties, thus allowing us to further discuss the epistemic role of Basic Evaluation Equations. The second issue relates to the classification of properties, or of property evaluations, in terms of *scale types*, and thus particularly to the characterization of quantities as specific kinds of properties, thus leading to the question whether, and under what conditions, non-quantitative properties can be measured. On this basis, the third problem is explored: the *conditions of existence of general properties* and the role of measurement in the definition of general properties.

6.1 Introduction

We have proposed that properties of objects are associated with modes of empirical interaction of objects with their environments, with the acknowledgment that this interaction makes objects experimentally comparable with one another. Thus we ground our framework upon the assumption that *through properties we account for the relational behavior of objects*. As already mentioned, this does not mean that we consider a property to exist only if an interaction is observed: our position is simply that observed interactions among objects can be accounted for in terms of their properties. We also accept that the description of an interaction among objects in terms of given properties is always revisable: there must be properties there, but they are not necessarily as we describe them.

As presented in the previous chapter, the framework we are developing is grounded on *individual properties*, such as lengths and reading comprehension abilities, which we take to be universal entities (see Sect. 5.3.2) that can be instantiated by, or more generically identified as, properties of given objects, such as the lengths of given rods and the reading comprehension abilities of given individuals.[1] A basic relation of the framework is then

a property of a given object identifies an individual property

or more shortly

a property of an object is an individual property

so that, for example, there is a length that a given rod has, i.e., the length of that rod (the property of an object) is that length (an individual property), and there is a reading comprehension ability that a given individual has, i.e., the reading comprehension ability of that individual (the property of an object) is

[1] In the case of quantities, it might be that individual quantities are those entities sometimes called “magnitudes”. However, the concept <magnitude> is used in radically different ways: quantities *are* magnitudes but also *have* magnitudes, as in the current edition of the VIM, which defines <quantity> as follows: “property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference” (JCGM, 2012: 1.1) (see the more extensive discussion in Footnote 19 of Chap. 2). Given this confusion, and the fact that measurement results can be, and usually are, reported without reference to magnitudes, we avoid including <magnitude> in the ontology we are presenting here (for an analysis of the relations between <quantity> and <magnitude> see Mari & Giordani, 2012: p. 761–763).

that reading comprehension ability (an individual property).

Each property of an object identifies an individual property: individual properties can be handled mathematically, for example by checking which of two lengths is greater, whereas relations between properties of objects must be investigated empirically. Accordingly, when a property of an object appears in a formal relation, such as a mathematical equation or a logical inference, the actual reference is to the corresponding individual property. This applies in particular to the relation of indistinguishability of properties of objects: as already pointed out, the observation that two properties of objects, $P[a_i]$ and $P[a_j]$, are indistinguishable, $P[a_i] \approx P[a_j]$, is interpreted by assuming that $P[a_i]$ and $P[a_j]$ either identify the same individual property or identify distinct but empirically indistinguishable individual properties. Since in general it is not possible to ascertain which of these situations is true, the customary notation $P[a_i] = P[a_j]$ is just a convenient shorthand, acceptable whenever the relation is assumed to be transitive (see Sect. 5.2.6 on the implications of this assumption).

As discussed in Sect. 2.2.3, comparable individual properties are said to be *of the same kind* (JCGM, 2012: 1.2). Kinds of properties are abstract entities that are reified by assuming the existence of corresponding *general properties*, so that the adjectives "long", "heavy", etc. are replaced by the nouns "length", "weight", etc., and a relation such as

$$long[a] \approx long[b]$$

as in Sect. 5.2.6, is more customarily written

$$length[a] \approx length[b]$$

Each individual property is then an instance of a general property, and two individual properties are comparable only if they are instances of the same general property. Again, the examples are obvious: any given length is an instance of length, any given reading comprehension ability is an instance of reading comprehension ability, and so on. A second relation of the framework is then

an individual property is an instance of a general property

These relations are depicted in Fig. 6.1.

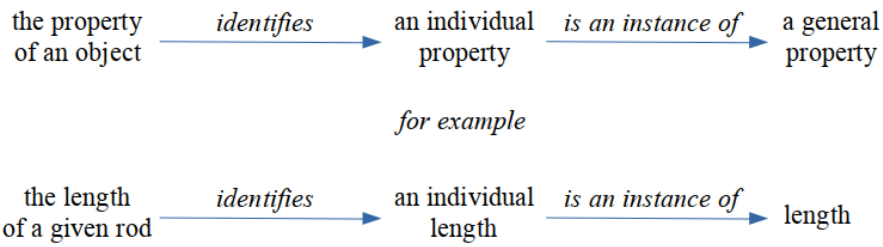


Fig. 6.1 Relations between properties of objects, individual properties, and general properties

While such a conceptualization might appear redundant, it is not hard to show that:

- properties of objects are not identical to individual properties: properties of objects are in fact features of objects and as such have a spatiotemporal location and can be (individual) measurands, and neither individual properties nor general properties share this feature; however, some features, such as being comparable with respect to a given relation, are characteristic of individual properties and are inherited by properties of objects; for example, we can say that the individual length ℓ_1 is greater than the individual length ℓ_2 if they are identified as the length of rod *a* and of rod *b* respectively and rod *a* has been empirically discovered to be longer than rod *b*;

- individual properties are not identical to general properties: individual properties can be comparable with each other, and general properties do not share this feature; however, some features, such as being a quantitative or a qualitative property, being a physical property or a psychosocial property, etc., are characteristic of general properties and are inherited by their instances; for example, a given length is a physical quantity because length is a physical quantity.

This provides a pragmatic justification of the structure illustrated in Fig. 6.1.[2]

On this basis, we defer to Sects. 6.5 and 6.6 a more specific analysis about general properties, in particular about their categorization into types, such as nominal, ordinal, and so forth. In the sections that follow, we continue to develop this framework grounded on individual properties, by introducing values of properties and first focusing on values of quantities.

6.2 Towards values of properties

A Basic Evaluation Equation, in its simplest version in which uncertainty is not taken into account, is

property of an object = value of a property

When Norman Campbell famously stated that "the object of measurement is to enable the powerful weapon of mathematical analysis to be applied to the subject matter of science" (1920: p. 267), it is plausible that he was indeed referring to this kind of equation, and expressly to the specific case

quantity of an object = value of a quantity

which, when written in the *Q*-notation (see Sect. 5.1), enables "the powerful weapon of mathematical analysis" by explicitly including numbers in the equation, e.g.,

$L[a] = 1.2345$ m

as multipliers of units (henceforth we write "$L[a]$" as a shorthand for "*length*[*rod a*]"). Analogous is the case of the modified notation

$L_{in\ metres}[a] = 1.2345$

as in some formalizations, such as those adopted in representational theories of measurement (see, e.g., Krantz et al., 1971, and also Kyburg, 1984: p. 17). Through Basic Evaluation Equations, values of properties, and thus values of quantities in particular, are indeed the mathematical counterparts of empirical properties of objects. Values play the fundamental role of providing the information that is the reason for which measurement is performed: before measurement the measurand is known only as a property of an object; after measurement we also know a value for it. (Once again, references to uncertainty are important for a more complete presentation of measurement, but are not relevant here.) Once the relation is accepted as dependable, the value can be mathematically manipulated in place of experimentally operating on the property of the object. As a trivial example, if $L[a] = 1.2345$ m and $L[b] = 2.3456$ m then from the mathematical fact that 1.2345 m < 2.3456 m we can immediately infer that $L[a] < L[b]$.

An analysis of the nature and the role of values of properties is then a core component for the development of a measurement-related ontology and epistemology of properties. Let us start by

[2] A foundational ontology might endeavor to build a framework on properties eventually based on one entity, from which everything else can be derived (an example of this monism is trope theory; see Maurin, 2018) or reduced, as nominalism would do by assuming that both individual properties and general properties are just concepts, and only properties of objects exist outside our minds (see Sect. 5.3.1). However, this philosophical task has no direct consequences for measurement science.

considering the specific case of quantities and their values.[3] (Henceforth we occasionally use the short term "value", rather than "value of a property" or "value of a quantity", if this does not create ambiguity.)

6.2.1 Values of properties: what they are not

Values of properties have such a critical role that it is perhaps not surprising that there are multiple and even incompatible positions on what they are. According to two common stereotypes, they are *expressions*, or they are *symbols*. Let us start our analysis by showing that neither of these positions is correct. These stereotypes are usually related to quantitative properties rather than properties as such: hence, *pars pro toto*, we refer to quantities in the discussion that follows.

First, for example, according to the first edition of the VIM the value of a quantity is "the *expression* of a quantity in terms of a number and an appropriate unit of measurement" (ISO, 1984: 1.17, emphasis added). The first definition that the Oxford English Dictionary (OED) gives of <expression> is "things that people say, write or do in order to show their feelings, opinions and ideas": thus, in general usage, according to the OED expressions (including mathematical expressions) are linguistic entities, i.e., in the sense of terminology, neither concepts nor objects (see Sect. 2.1). But it should be clear that here, in discussing measurement, values are not linguistic entities. Consider the difference, e.g., between the rod *a*, which is an object, and the five-character (space included) term "rod *a*", which is a linguistic entity: the object has a weight, a color, etc., whereas the term does not. The term "rod *a*" *refers* to a given rod, but *is* not that rod. Analogously, values are communicated by means of terms but they are not terms.[4] And in fact the same value, e.g., 1.2345 m, can be expressed linguistically in multiple ways, e.g., "one point two … metres", "1.2345 m", and "1,2345 m" (for most non-English speaking people), showing that 1.2345 m and "1.2345 m" are different entities. Certainly, values must somehow be expressed by means of linguistic entities to be communicated, but they are not, in themselves, expressions.

Second, sometimes values are said to be *symbols*, or *identifiers*, which stand for or represent objects or quantities of objects.[5] Of course, values may well be *used* as such, but this does not solve the problem of what they *are*. Indeed, stating that *x* is a symbol of *y* does not say anything about what *x* is. In this sense, Napoleon can be a symbol of political power, and a sphere can be a symbol of perfection, but this does not change the fact that Napoleon was a human being and a sphere is a geometric object. "To be a symbol" is just convenient shorthand for "to be used as a symbol". Hence values may be used as symbols to represent quantities of objects, but a definition of <value of a quantity> phrased as "symbol such that..." is ontologically vacuous.

[3] In fact, the analysis that follows may be easily generalized to the case of properties and values of properties, as we do later on in this chapter, where we also discuss the characterization of quantities as specific kinds of properties. We start by presenting the more specific case of values of quantities because the very concept <value of a property> is not widely used, and some would consider it controversial. It should be noted that the boundary between quantitative and non-quantitative properties is not uniquely defined, and in particular there are controversies whether ordinal properties are quantitative or not. However, that additive properties are quantities is not an issue, and we start our discussion from them.

[4] One example of a term used to communicate a value is a *numeral*, which is a term for a number; for example, "4" and "IV" are both numerals that stand for the number 4. As was previously discussed, Campbell (1920) defined measurement as "the assignment of numerals to represent properties", and Stevens (1959) defined measurement as "the assignment of numerals to objects or events according to rule"; such statements may have inadvertently contributed to the confusion between values and terms. It may be worth noting that even though both Campbell and Stevens are both associated with representational theories of measurement, the wider literature on representationalism emphasizes the mapping of objects, or possibly properties, to numbers, not numerals, as discussed further below.

[5] For example, André Weyl wrote that "measurement permits things … to be represented conceptually, by means of symbols" (1949: p. 144). While not false, this claim is by no means characteristic of measurement in particular, and therefore is not very informative.

6.2.2 Values of properties cannot be discarded in contemporary measurement

At this point we need to face the possible objection that values are not needed at all, and therefore our whole current problem can be dismissed as immaterial. At least two analogous arguments can be made in support of this position.

One argument is that most equations and the related explanations that appear in the literature on, for example, physics do not even mention units: while often introduced as relations among general quantities (e.g., $F = ma$), physical laws are also interpreted as equations that relate numerical values of such quantities, under the assumption that their units are consistently chosen in a system of units. Hence it would seem that, after a system of units has been chosen, values can be discarded, and instead one need only to report numbers, instead of values (e.g., 1.2345 instead of 1.2345 m), for conveying information about quantities of objects.

The second argument against values of properties starts from the supposition that measurement produces numbers rather than values. As mentioned above, this seems to be assumed in particular by representational theories of measurement (see, e.g., Krantz et al., 1971), which usually formalize measurement as a mapping from objects, or sometimes properties of objects (see also Sect. 5.2.5), to numbers[6] by maintaining the unit implicit in the mapping, thus re-writing, e.g., $L[a]$ = 1.2345 m as $L_{in\_metres}[a]$ = 1.2345. This seems to be a reinterpretation of Russell's well-known assertion that "Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real, as the case may be" (1903: p. 176). Indeed, the Q-notation (see Sect. 5.1)

$$Q[a] = \{Q[a]\}\ [Q]$$

is equivalent to

$$Q[a] / [Q] = \{Q[a]\}$$

where then $L[a]$ / m is what $L_{in\_metres}[a]$ is actually meant to be. Since in this relation values of quantities seem to have disappeared, it might be concluded that they are only related to the way knowledge is represented and therefore that they can be avoided by an appropriate choice of the representation.

As we see them, both of these arguments are correct in their premises, but their conclusions are problematic: the fact that in specific cases values can actually be discarded, in favor of dealing with numbers only, is really just a sort of shorthand and does not imply that this is always the case. Rather, there are good reasons for the customary choice of writing the Basic Evaluation Equation in terms of values instead of numbers. The difference between values of quantities and numerical values is that only the former contain information on the metrological context: "1.2345 m" means <1.2345 in the context of the scale generated by the metre>. Reporting only a numerical value, such as 1.2345, loses

[6] The fact that distinct objects can have the same quantity, e.g., the same length, and therefore are mapped to the same number, makes the quantity-related mapping non-injective, thus a homomorphism. What Louis Narens wrote (1985: p. 7) on this matter is interesting (note that he uses the term "scale" to refer to such mappings): "I often prefer to change the character of the representational theory a little and consider a scale to be an isomorphism between the empirical or qualitative situation and some mathematical situation. The primary reason for this is that isomorphisms preserve truth whereas homomorphisms do not.". According to the ontology we are proposing, a way for making the mapping injective, and therefore an isomorphism, is to assume that its domain is the set of individual properties, rather than of the properties of objects or of objects. Our ontology highlights that individual properties can be measured only in their being properties of objects, thus making the mapping that formalizes such an experimental process non-injective. (Admittedly, consistently with this thinking, Narens chose to title his book "*Abstract* measurement theory" (emphasis added); hence perhaps the prior question is whether the very concept <abstract measurement> has anything to do with actual measurement as it is commonly understood.)

the reference to such a context, which is crucial for guaranteeing the metrological traceability of measurement data.

Assertions such as Russell's hide the issue by implicitly assuming that the metrological context is given and is entirely embedded in the definition of the general quantity under measurement, as if a "natural unit of length" were unproblematically available, allowing us to measure the "natural length" of any object by a number, interpreted as the multiple of such a "natural unit" and conveying the information of the traceability to such a unit. It is in fact as if measurement could always be, in its structure, the counting of "natural units".

But unless and until such "natural units" for all relevant quantities are agreed upon and socially accepted,[7] it is convenient, and essential, for Basic Evaluation Equations, and measurement results, to contain information on their metrological context, as provided by values, which thus play a critical role in effective communication of the information acquired by means of measurement.

On this basis, let us continue our exploration of what values of quantities are.

6.3 Constructing values of quantities

While the concept <value of a property> might appear unusual (as an example, the VIM does not define it), values of *quantities* are widely used, uncontroversially recognized as multiples of units. Even those who are doubtful about the nature of values of quantities, as discussed above, accept that 1.2345 m and 2.34 kg are examples of them. In order to properly introduce values of properties in our framework, let us then start from values of quantities, by exploiting the familiar additive structure of quantities such as length. What follows is a construction by example, rather than a definition.

6.3.1 Operating on (additive) quantities of objects

Let us consider two rods, r and r', in the experimental situation depicted in Fig. 6.2.

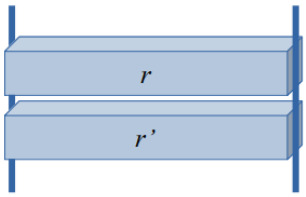


Fig. 6.2 Constructing values of quantities: first step (quantity-related comparison)

This situation is usually described as

the rods r and r' have the same length

or

the length of r is the same as the length of r'

and therefore

$$L[r] \approx L[r']$$

[7] This highlights another barrier to the elimination of values of quantities in favor of numbers: for all properties evaluated in scales of types algebraically weaker than ratio (see the related discussion in Sect. 6.5), the social acceptance of "natural units" is not sufficient. In particular, in the case of an interval scale a "natural zero" would also need to be universally adopted. The fact that Celsius and Fahrenheit scales of temperature have different zeros witnesses that this could not be a trivial task.

thus highlighting, more explicitly than $L[r] = L[r']$, that this is an experimental relation and therefore such a sameness is operationally a length-related indistinguishability.[8]

Moreover, let us then assume that, at least for objects such as rods, length is an empirically additive quantity,[9] so that there exists a length-related concatenation operation $\oplus$ (hence the symbol "$\oplus$" is used to denote an operation that applies to lengths of objects, not numbers) and the situation depicted in Fig. 6.3 is described as

> the length of a is indistinguishable from the length of the length-related concatenation of r and r'

or

$$L[a] \approx L[r] \oplus L[r']$$

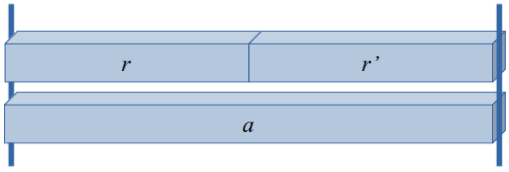


Fig. 6.3 Constructing values of quantities: second step (quantity-related concatenation)

Since $L[r] \approx L[r']$, this relation can also be written as

$$L[a] \approx L[r] \oplus L[r]$$

and therefore

$$L[a] \approx 2\, L[r]$$

for short, where more generally $n\, L[r]$, for any integer $n > 0$, denotes the length of n concatenated copies of $L[r]$.[10] This principle can be then extended also to non-integer relations between $L[a]$ and $L[r]$, by considering, together with *iteration*, $L[a] \approx n\, L[r]$, the inverse operation of *partition* (as the terms "iteration" and "partition" are used by Weyl, 1949: p. 30), such that $L[r]$ is assumed to be constituted of n' indistinguishable lengths $L[c]$, so that $L[r] \approx n'\, L[c]$. By combining the two operations, a length is obtained as $n/n'\, L[r]$. In varying the ratio n/n' a set of lengths is thus obtained, and while the construction starts from the length of a given object, r, each entity $n/n'\, L[r]$ is a length constructed without an object that bears it: what sort of entities are they, then? While leaving this question open for the moment, let us point out that *all these relations involve only quantities of objects*, and are obtained by experimentally comparing objects.

Suppose now that the length $L[r]$ is agreed to be taken as a *reference quantity* and given an identifier for convenience, say "ℓ_{ref}" (or, for example, "metre"). The reference length ℓ_{ref} is then *defined* as the length of the object r

[8] For the sake of simplicity, rods are modeled here as geometrically regular bodies, and in particular prisms, so that each rod has *one* length. Moreover, this construction is assumed to be performed in one inertial frame of reference, so that problems due to relativistic effects do not arise.

[9] We will relax this assumption later, in Sects. 6.3.6 and 6.3.7, in constructing values of less-than-ratio properties.

[10] The length $n\, L[r]$ is customarily defined by induction: $1\, L[r] := L[r]$, and $n\, L[r] := (n–1)\, L[r] \oplus L[r]$. Since we are operating with empirical quantities, not numbers, one might challenge the correctness of the equation $L[r] \oplus L[r] = 2\, L[r]$, contesting, in particular, that the geometry of our world on the one hand and the features of our instruments on the other hand do not allow us to guarantee the *perfect* collinear concatenation of rods. The argument is that numerically $L[a] \oplus L[b] = (L[a]^2 - 2\cos(\vartheta)\, L[a]\, L[b] + L[b]^2)^{½}$, where ϑ is the angle between the rods a and b, so that substituting $L[r] \oplus L[r]$ with $2\, L[r]$ is correct only if $\vartheta = \pi$, i.e., in the case of collinearity. This is true, of course, but the same argument can be exploited to provide an empirical check of collinearity, via the condition that $\oplus$ is associative: it is indeed trivially proved that for $(L[a] \oplus L[b]) \oplus L[c] = L[a] \oplus (L[b] \oplus L[c])$ to hold ϑ must be π (or, interestingly, $(1+2k)\pi/2$, for k=0, 1, ..., where the Pythagorean theorem applies: in a peculiar world, "collinear concatenation" means concatenation at right angles...).

$$\ell_{ref} := L[r]$$

and r can be called a *reference object*. The indistinguishability relation $L[a] \approx 2\ L[r]$ can then also be written as

$$L[a] \approx 2\ \ell_{ref}$$

This shows that the following relations

$$L[a] \approx L[r] \oplus L[r']$$

$$L[a] \approx L[r] \oplus L[r] \qquad \text{(provided that } L[r] \approx L[r']\text{)}$$

$$L[a] \approx 2\ L[r] \qquad \text{(a shorthand of the previous relation)}$$

$$L[a] \approx 2\ \ell_{ref} \qquad \text{(according to the definition of } \ell_{ref}\text{)}$$

all refer to the same empirical situation and only differ in the way the information is conveyed: in terms of the distinction between senses and referents of expressions (as explained in Sect. 5.3.2), the senses of the involved expressions are different, but their referent is always the same. All these relations – including the last one – involve lengths, and the difference between the length $L[r] \oplus L[r']$ and the length $2\ \ell_{ref}$ is only about how such lengths are identified.

The fact that this construction has been developed with no references to values is important. As Alfred Lodge noted (1888: p. 281)

> the fundamental equations of mechanics and physics express relations among quantities, and are independent of the mode of measurement of such quantities; much as one may say that two lengths are equal without inquiring whether they are going to be measured in feet or metres; and indeed, even though one may be measured in feet and the other in metres. Such a case is, of course, very simple, but in following out the idea, and applying it to other equations, we are led to the consideration of product and quotients of concrete quantities, and it is evident that there should be some general method of interpreting such products and quotients in a reasonable and simple manner.

With this acknowledgment we may restart our construction.

A rod a can be now calibrated in terms of its length with respect to ℓ_{ref} by aligning the left ends of a and r and placing a mark on the rod a at the other end of the rod r. Additional marks can be placed on the rod a, using geometrical methods that implement the iteration and partition methods mentioned above, to denote multiples of ℓ_{ref}, as depicted in Fig. 6.4.

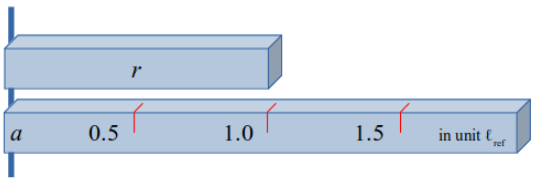


Fig. 6.4 Constructing values of quantities: third step (quantity-related comparison with an object calibrated with respect to a reference quantity)

Common measuring instruments of length, such as metre sticks and tape measures, are constructed and then calibrated in this way: indeed, the rod a can be placed against other objects to establish where their lengths align with corresponding marks on the rod a itself. Hence, the rod a realizes a sequence of lengths. The length $L[b]$ of an object b can be now compared with the lengths marked on the rod a and thus reported as a multiple x of ℓ_{ref},

$$L[b] \approx x\ \ell_{ref}$$

where then $x = n/n'$, for given n and n', as depicted in Fig. 6.5.

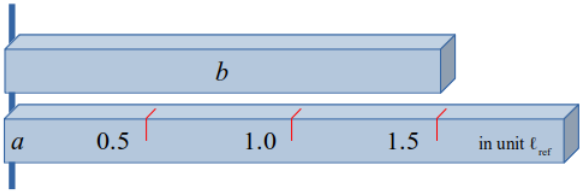


Fig. 6.5 The comparison of the length $L[b]$ with the lengths marked on the rod a

6.3.2 On reference objects and reference quantities

Let us now focus on the indistinguishability relation

$$L[b] \approx x\ \ell_{ref}$$

which holds for a given $x = n/n'$. The only difference between $L[b] \approx x\ \ell_{ref}$ and $L[b] \approx x\ L[r]$ is related to the way in which the quantity in the right-hand side of the two relations is referenced, by an identifier of the quantity, ℓ_{ref}, or by addressing an object, r, with respect to one of its properties, L. Since changing the way in which a quantity is referenced ("the length of the object we agreed to designate as r", "$L[r]$", "ℓ_{ref}", or whatever else) does not change the quantity, one might conclude that this is just an arbitrary lexical choice. While in principle this is correct, there is a subtle point here related to the way we usually deal with identifiers: for the relation (identifier, identified entity) to be useful, *it needs to hold in a stable way*. This is why entities whose time-variance is acknowledged are identified by means of identifiers indexed by a time-related variable, as in the case of the length $L[b, t]$ of the object b at the time t (see also Sect. 5.2.5). Conversely, if the identifier does not include a reference to time then the identification is supposed to remain valid only on the condition that the identified entity does not change over time. For example, the date of birth of a given person b can be identified as $birthday[b]$, while her height in a given time t as $height[b, t]$: in this way we acknowledge that birthday is time invariant, whereas height is time variant.

In this sense, the definition $\ell_{ref} := L[r]$, where the identifier "ℓ_{ref}" is not indexed with time, assumes that the length $L[r]$ is time invariant. Since quantities of objects are instead usually subject to variations, this is a strong assumption: of course, assigning a name to the quantity of an object does not make it stable.[11]

The consequence of choosing the length of an object r as a reference length ℓ_{ref}, thus under the condition of its stability, is that ℓ_{ref} can also be considered to be the length of any other sufficiently stable object having the same length as r. This allows the assessment of $L[a] \approx x\ \ell_{ref}$ not only by means of $L[a] \approx x\ L[r]$ but also by means of $L[a] \approx x\ L[r']$, for any sufficiently stable r' in a class of objects such that $L[r'] \approx L[r]$. Hence the choice of referring to a length through an identifier as "ℓ_{ref}" (for example "metre" – note: it is not "metre in a given time t") assumes that the referenced length is both space and time invariant: according to the conceptual framework introduced in Sect. 6.1, it is an individual length, identified by $L[r]$, $L[r']$, ... but abstracted from any particular object.[12]

[11] This problem is arguably even more pernicious in the human sciences, wherein properties commonly vary not only by time but also by socio-cultural-historical context, as also discussed in Sect. 4.4.

[12] The assumption that, though initially defined about an object, a reference quantity is an abstract, individual quantity is what justifies the notation "ℓ_{ref}", with the symbol for the quantity written in lowercase roman (see Table 2.3 and Footnote 24 of Chap. 2).

6.3.3 Alternative reference quantities and their relations, i.e., scale transformations

As remarked, the only condition for having singled out r as a reference object is that its length is stable. Hence nothing precludes the independent choice of an alternative reference object, r^*, whose length $L[r^*]$ is distinguishable from $L[r]$ and defines a new reference length (for example the foot instead of the metre):

$$\ell_{ref^*} := L[r^*]$$

A new rod a^* can be now calibrated with respect to ℓ_{ref^*}, exactly as was done before for the rod a with respect to ℓ_{ref}, so that the same object b could be compared in its length with both rod a and rod a^*. Different relations of indistinguishability are then obtained, $L[b] \approx x\ \ell_{ref}$ and $L[b] \approx x'\ \ell_{ref^*}$, with $x \neq x'$, as exemplified in Fig. 6.6.

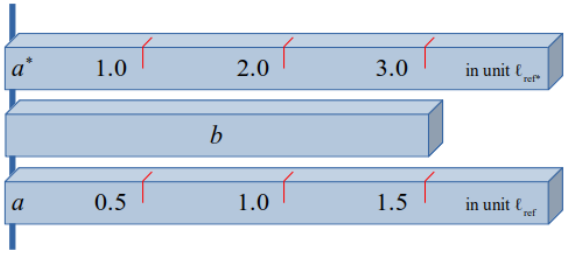


Fig. 6.6 The comparison of the length $L[b]$ with the lengths marked on the rods a and a^*

The lengths marked in this way on rods a and a^* can be compared, which is particularly interesting because such lengths are indexed by numbers, attributed according to the hypothesis of empirical additivity, such that the length 2 ℓ_{ref} is $L[r] \oplus L[r]$ and so on. Hence, the hypothesis that the lengths marked on two rods have been additively constructed *can be experimentally validated*, by finding the factor k such that $\ell_{ref} \approx k\ \ell_{ref^*}$ (in the example in Fig. 6.6, $k = 0.5$) and then checking whether 2 $\ell_{ref} \approx 2k$ ℓ_{ref^*}, 3 $\ell_{ref} \approx 3k$ ℓ_{ref^*}, and so on. Such a systematic validation provides a justification for the specific hypothesis that the two lengths ℓ_{ref} and k ℓ_{ref^*} are in fact equal, $\ell_{ref} = k\ \ell_{ref^*}$, and not just indistinguishable, and therefore that the *scale transformation*, from multiples of ℓ_{ref} to multiples of ℓ_{ref^*} or vice versa, can be performed as a mathematical operation.[13]

6.3.4 Generalizing the strategy of definition of reference quantities

The definition of reference quantities as quantities of objects (sometimes called "prototypes" or "artifacts" when they are physical objects) that are hypothesized to be stable is conceptually simple, and is typically the starting point of the development of a unit. For example, in 1889 the first General Conference of Weights and Measures (CGPM) asserted that "the Prototype of the metre chosen by the CIPM [...] at the temperature of melting ice shall henceforth represent the metric unit of length." (BIPM, 2019: Appendix 1), where the mentioned prototype of the metre was a specially manufactured metallic rod. But this strategy has some drawbacks that have become more and more apparent with the progressive globalization of measurement science and its applications:

- first, both physical and non-physical objects at the anthropometric scale are usually not completely stable, with the consequence that, once the definition $\ell_{ref} := L[r]$ is given, for any object a if $L[a] \approx x\ \ell_{ref}$ and $L[r]$ changes due to the instability of r, then after that change $L[a]$

[13] The inverse approach is also possible: given a predefined reference length ℓ_{ref} and a given factor k, a new reference length ℓ_{ref^*} could be defined as $\ell_{ref^*} := k\ \ell_{ref}$. In this case, finding an object r^* such that $L[r^*] = \ell_{ref^*}$ (thus an empirical relation, not a definition) would correspond to realizing the definition of the new reference length.

$\approx x' \, \ell_{\text{ref}}$, with $x' \neq x$, even if $L[a]$ did not change: the numerical representation of a quantity has changed even though the quantity itself did not;[14]

- second, having a reference quantity defined as the quantity of one object implies that all traceability chains must start from that object, in the sense that all measuring instruments for that quantity must be directly or indirectly calibrated against that reference object: this is operationally inconvenient and may generate political struggles, given the power that the situation confers to the owner of the object.

Alternative strategies to define stable and accessible reference quantities may be and have been in fact envisaged to avoid or at least reduce these flaws. Such alternative strategies are particularly required in the case where the quantities intended to be measured are properties of human beings, which, if the steps described above were to be followed, would imply that, in principle, reference objects should be certain individual humans, a situation that of course is not usually appropriate for several reasons.[15]

Rather than selecting specific objects, a representative sample of objects – hence persons, in some cases – could be then selected, their property of interest somehow evaluated, and the reference quantity defined as a statistic (e.g., the mean or the median) of the obtained values. This makes the reference quantity depend on the selected sample, and therefore in principle it is not entirely stable if new samples are taken or if characteristics of the sampled population change. (In psychometrics, evaluations performed according to this strategy are called "norm-referenced", to emphasize the normative role of the sample that defines the reference quantity; see Glaser, 1963.[16])

Another possible strategy for dealing with these issues may be based on the consideration that according to the best available theories there is a class of objects, $R = \{r_i\}$, that when put in given conditions invariantly have the same quantity of interest, which in those conditions is then a constant.[17] Defining the reference quantity as a constant quantity characteristic of a class of objects guarantees both its stability and accessibility. And if the identified constant were too far from the anthropometric scale to be suitable, the reference quantity could be defined as an appropriate multiple or submultiple of the constant, so as to maintain a principle of continuity, such that different definitions could subsequently be adopted while ensuring that the defined reference quantity remains the same. For example, in 1960 the 11th General Conference of Weights and Measures redefined the metre as "the

[14] The fact that this is possible is a compelling reason to maintain the distinction between the quantities in the left- and right-hand sides of Basic Evaluation Equations. Unfortunately, this is sometimes confused. Take the following example: "Suppose we had chosen as our standard [of mass] a cube of iron rather than platinum. Then, as the iron rusted, all other objects would become lighter in weight" (Kaplan, 1964: p. 186). This is wrong: the other objects do *not* become lighter, they only seem to be lighter when compared to the rusted cube, given that it is only the numerical representation of their mass that changes. A very well-studied case of such changes is that of the kilogram, which before the 2019 revision of the SI was defined as the mass of a given artifact, the International Prototype of the Kilogram (IPK). In 1994 a periodic verification of several national copies of the prototype of the kilogram revealed that basically all of them appeared as if they had gained some mass in comparison with the IPK, despite their independent handling and storage: plausibly, it was instead the IPK that has lost mass (Girard, 1994).

[15] For discussions of strategies for defining reference quantities in human measurement using resources from the Rasch measurement tradition, see, e.g., Maul et al. (2019), Wilson et al. (2019), and Briggs (2019).

[16] An example of this is the well-known case of the intelligence quotient (IQ), defined by taking the median raw score of the chosen sample as IQ 100 and one sample standard deviation as corresponding to 15 IQ points. It has been observed that since the early 20th century raw scores on IQ tests have increased in most parts of the world, a situation called the *Flynn effect* (Flynn, 2009). Whether intelligence as such has also increased is, of course, another matter.

[17] On this matter Eran Tal (2019) refers this situation to the distinction between *types* and *tokens* (see Wetzel, 2018), and proposes the thermal expansion coefficient of aluminum at 20 °C and the thermal expansion coefficient of a particular piece of aluminum at a given temperature as examples of a type and a corresponding token respectively, such that "quantity types may be instantiated by more than one object or event". Since the thermal expansion coefficient of aluminum at 20 °C is a given thermal expansion coefficient, which is an individual quantity, a quantity type, in Tal's lexicon, can be thought of as a "partially abstracted" individual quantity, while any non further specified given thermal expansion coefficient is then a "fully abstracted" individual quantity.

length equal to 1 650 763.73 wavelengths in vacuum of the radiation corresponding to the transition between the levels $2p_{10}$ and $5d_5$ of the krypton 86 atom" (BIPM, 2019: Appendix 1). The critical point of this definition is the assumption that the wavelength of the chosen radiation is constant, whereas the numerical value, 1 650 763.73, was only chosen for the purpose of guaranteeing that the metre remained the same length despite the change of its definition.[18]

By exploiting the functional relations that are known to hold among quantities, a more sophisticated version of this strategy allows for the definition of a reference quantity as a function of constants of different kinds, and possibly of previously defined reference quantities. For example, according to Einstein's theory of relativity the speed of light in vacuum is constant, and so the class of all light beams in vacuum is such that the length of their path in a given time interval is also constant. By exploiting the relation

length = speed x time duration

among general quantities, the definition is then

$$\ell_{\text{ref}} := S[\text{R}]\,\Delta T$$

where S[R] is the speed S of light in vacuum (R being then intended as the class of all light beams in vacuum) and ΔT is the chosen time interval. This is in fact how in 1983 the 17th General Conference of Weights and Measures defined the metre: "the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second" (BIPM, 2019: Appendix 1). Once again, the appropriate choice of the numerical value, 1/299 792 458, was the condition of validity of the principle of continuity.

With the aim of emphasizing the role of the defining constant quantity S[R], this definition can be rephrased as

the reference length ℓ_{ref} is such that $S[\text{R}] = \ell_{\text{ref}}\,\Delta T^{-1}$

and this is in fact what became the definition of the metre in 2019 as a result of the 26th General Conference of Weights and Measures: "The metre (...) is defined by taking the fixed numerical value of the speed of light in vacuum c to be 299 792 458 when expressed in the unit m s^{-1}, where the second is defined in terms of the caesium frequency $\Delta\nu_{\text{Cs}}$." (BIPM, 2019: 2.3.1).

Given the condition of the correctness of the theory that considers a quantity constant for a class of objects, this generalization produces three important benefits, by making the unit

- independent of the conditions of stability of a single object,
- more widely accessible (in principle, everyone with the access to one object of the class can realize the definition of the unit, and therefore operate as the root of a traceability chain), and
- definable in terms of quantities of kinds other than that of the unit, given the condition that all relevant quantities are related in a system of quantities.

This developmental path, where a unit is defined as

1. the quantity of a given object (*prototype-based* definition), then
2. a statistic of the quantities of a representative set of objects (*norm-referenced* definition), then
3. the quantity considered to be constant for a class of objects (*constant-based* definition, as in the 1960 definition of the metre), then
4. a quantity functionally related to the quantity/ies considered to be constant for a class of

[18] This is one more case in which the distinction between sense and reference (see Sect. 5.3.2) is relevant. The assumption of validity of the principle of continuity can be written as $\text{metre}_{1889} = \text{metre}_{1960}$, in which the fact that the metre was defined in different ways in 1889 and in 1960 makes the senses of the two expressions ("the metre as defined in 1889" and "the metre as defined in 1960") different, while their referents are the same.

objects (*functional constant-based* definition, as in the 1983 definition of the metre), with the relation possibly stated in inverse form (*inverse functional constant-based* definition, as in the 2019 definition of the metre),

may be interpreted as a blueprint of the options for the definition of reference quantities: in fact, a lesson learned from history.

6.3.5 Values of quantities: what they are

Let us summarize the main features of the construction proposed in the previous sections. In the special case of an empirically additive general quantity Q, the quantities $Q[a_i]$ of objects a_i can be concatenated so that the concatenation $Q[a_i] \oplus Q[a_j]$ can be empirically indistinguishable from a quantity $Q[a_k]$, that is, $Q[a_i] \oplus Q[a_j] \approx Q[a_k]$.[19] On this basis an object r having the quantity Q can be singled out with the conditions that it is sufficiently Q-stable and that Q-related copies of it are available. This allows for the identification of the individual quantity $Q[r]$ not only as "$Q[r]$" – i.e., the quantity Q of the object r – but also through a time-independent identifier "$\mathrm{q}_{\mathrm{ref}}$" ("$\ell_{\mathrm{ref}}$" in the example above). This also allows for reporting of the information on a quantity $Q[a_i]$ in terms of its indistinguishability from a multiple x of $\mathrm{q}_{\mathrm{ref}}$, $Q[a_i] \approx x\ \mathrm{q}_{\mathrm{ref}}$. Furthermore, other such reference objects r^* can be chosen, and the scale transformation $\mathrm{q}_{\mathrm{ref}} = k\ \mathrm{q}_{\mathrm{ref}^*}$ can be experimentally tested, for a given k that depends on $\mathrm{q}_{\mathrm{ref}}$ and $\mathrm{q}_{\mathrm{ref}^*}$.

While everything that has been done in this construction is related to quantities of objects, the conclusions apply to what are commonly acknowledged to be *values* of quantities, and in fact the indistinguishability

$$Q[a_i] \approx x\ \mathrm{q}_{\mathrm{ref}}$$

can be interpreted as a Basic Evaluation Equation

quantity of an object ≈ value of a quantity

as well, as follows:

- an individual quantity $\mathrm{q}_{\mathrm{ref}}$ is singled out as a quantity unit (e.g., the metre); $\mathrm{q}_{\mathrm{ref}}$ may be *defined* as the quantity $Q[r]$ of an object r or, being defined in some other way as discussed in Sect. 6.3.4, may be *realized* by some object r; in either case r is a measurement standard, and possibly in particular the/a *primary standard*;
- the individual quantities $x\ \mathrm{q}_{\mathrm{ref}}$ (e.g., 2 m) are values of quantities, being by construction, the multiples of $\mathrm{q}_{\mathrm{ref}}$ obtained by means of the concatenation of the chosen unit;
- working standards r' can be calibrated against the primary standard r, $Q[r'] \approx Q[r]$ (ignoring calibration uncertainty), so that the quantity $Q[a]$ of an object a can be compared with $Q[r']$; hence the inference that from $\mathrm{q}_{\mathrm{ref}} = Q[r]$, $Q[r'] \approx Q[r]$, and $Q[a] \approx x\ Q[r']$ leads by transitivity[20] to $Q[a] \approx x\ \mathrm{q}_{\mathrm{ref}}$ is the simplest case of a *metrological traceability chain* (JCGM, 2012: 2.42);
- the relation $Q[a] \approx x\ \mathrm{q}_{\mathrm{ref}}$ for a given x (e.g., $L[a] \approx 2$ m) is a Basic Evaluation Equation, and thanks to this traceability it may be a measurement result for the measurand $Q[a]$ (ignoring measurement uncertainty);
- hence the relations

[19] A generalized version of this condition is usually part of an axiomatic system of quantities. For example, the seventh axiom of Patrick Suppes' system (1951: p. 165) is, in our notation: if $Q[a_i] \leq Q[a_k]$ then there exists a number x such that $Q[a_k] = x\ Q[a_i]$.

[20] In Sect. 5.2.6 we pointed out that indistinguishability is generally not transitive: how traceability chains can be constructed in spite of this obstacle is discussed by Mari and Sartori (2007).

the quantity of a given object is indistinguishable from a multiple of the quantity of another object

(e.g., $L[a] \approx 2\ L[r]$) and

the quantity of a given object is a value of a quantity

(e.g., $L[a] \approx 2\ \ell_{ref}$ (or $L[a] \approx 2$ m, that indeed is commonly read "the length of *a is* 2 metres")) refer to the same empirical situation, the difference being in the way the two relations convey the information about the individual quantities involved.

The conclusion is then obvious: *a value of a quantity is an individual quantity identified as a multiple of a given reference quantity, designated as the unit*.[21]

The analysis in Sect. 5.3.2, which led us to interpret a relation such as $Q[a_i] \approx Q[a_j]$ as including expressions with different senses but possibly the same individual length as their referent, can be now straightforwardly extended to scale transformations and Basic Evaluation Equations:

- in the scale transformation $q_{ref} = k\ q_{ref*}$ the expressions "q_{ref}" and "$k\ q_{ref*}$" have different senses but the same individual length as referent;[22]
- in the Basic Evaluation Equation $Q[a] \approx x\ q_{ref}$, for a given x, the expressions "$Q[a]$" and "$x\ q_{ref}$" have different senses but could have the same individual length as their referent.

The concept system about <quantity> can then be depicted as in Fig. 6.7.

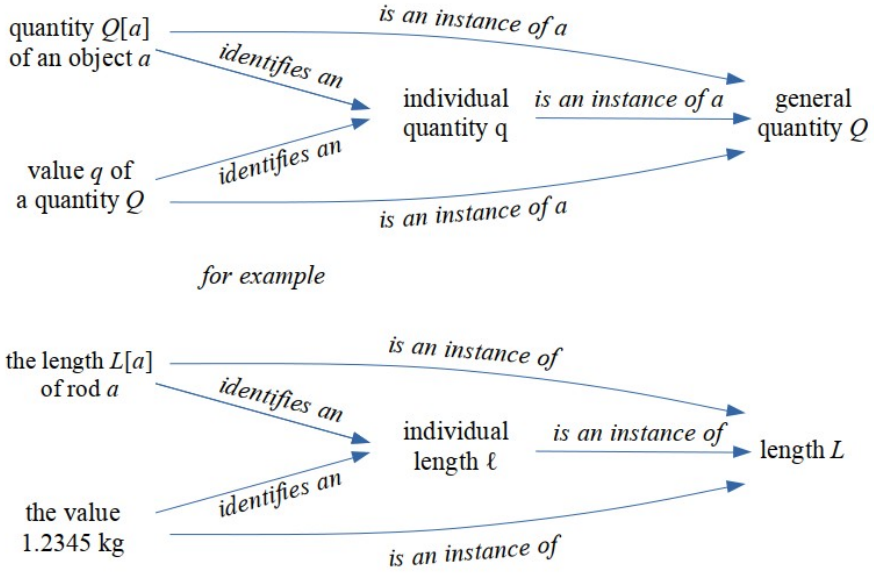


Fig. 6.7 The concept system about <quantity> (top) and an example (bottom), just as a specialization of Fig. 5.2

As discussed in Sect. 5.1, there is nothing arbitrary in the fact that an individual quantity q is identified as the quantity $Q[a]$ of an object a. Once again, this shows that the Basic Evaluation Equation $L[a] =$ 1.2345 m conveys the information that there is an individual length ℓ such that both the length $L[a]$ of rod a and the value 1.2345 m are claimed to be instances of ℓ. This allows us to propose a short

[21] Of course, we are still referring here only to quantities with a unit, and not, in particular, to ordinal quantities, that are discussed later on. As explained in Footnote 12 of Chap. 5, we use the concept <multiple of a quantity> in a broad sense, admitting also non-integer multiples.

[22] As a consequence, we can provide a simple answer to a question such as whether, e.g., 1.2345 metres and 48.602 inches are the same value or not: "1.2345 m" and "48.602 in" have the same referent – i.e., 1.2345 metres and 48.602 inches are the same length – but they have different senses. For short, they are *conceptually different* but *referentially the same*. Whether this leads to the conclusion that they are the same value or not depends on the specific definition of <value> that is adopted.

discussion, in Box 6.1, about the delicate subject of true values of quantities.

Box 6.1 – True values of quantities

Plausibly due to its explicit reference to truth, the idea that the quantity of an object may have a true value, which measurement aims at discovering, has caused controversies and confusion in measurement science for decades, if not centuries. It is a position that has appeared so deeply entangled with philosophical presuppositions about the actual existence of an observer-independent reality that a mutual understanding seemed impossible without preliminary agreement about an underlying ontology. Consider two authoritative, traditional examples of positions that jointly illustrate this controversy. According to Ernest Doebelin "when we measure some physical quantity with an instrument and obtain a numerical value, usually we are concerned with how close this value may be to the 'true' value" (1966). Vice versa, Churchill Eisenhart wrote about his "hope that the traditional term 'true value' will be discarded in measurement theory and practice, and replaced by some more appropriate term such as 'target value' that conveys the idea of being the value that one would like to obtain for the purpose in hand, without any implication that it is some sort of permanent constant preexisting and transcending any use that we may have for it." (1962).

While acknowledging that references to truth are usually laden with philosophical presuppositions, our understanding of what values of properties are and of the epistemic role of the Basic Evaluation Equation allows us to propose a simple interpretation of <true value>.

As a preliminary note, it should be clear that truth or falsehood do not apply, properly, to values: asking whether, say, 1.2345 m is true or false is meaningless. Rather, the claim that 1.2345 m is a true value refers to its relation with a property of an object, say, the length of rod a, being thus just a shorthand for the assertion that the Basic Evaluation Equation

$$L[a] = 1.2345 \text{ m}$$

is true. Of course, there can be several reasons that may make it hard or even impossible to achieve knowledge of the actual truth or falsehood of this equation, and the equation as such could be ill defined.

These issues may be left aside in a first, principled analysis about <true value>, which maintains in particular the distinction between the *existence* of true values and the possibility of our *knowledge* of them. Indeed, the VIM notes that true values are "in principle and in practice, unknowable" (JCGM, 2012: 2.11, Note 1), but this may be simply considered as an instance of the recognition that we cannot be definitively certain of anything related to empirical facts, like values of empirical properties of objects.

If true values are interpreted in the context of Basic Evaluation Equations, the conclusion does not seem to be particularly controversial: as already discussed, in Sect. 5.1.3 and elsewhere, asking whether 1.2345 m is the true value of the length of rod *a* is the question of whether there is an individual length that is at the same time the length of rod *a* and 1.2345 times the metre. Moreover, if values of properties and evaluation scales are considered as classification tools, as further discussed also in Box 6.2, the truth of a Basic Evaluation Equation is about the correctness of the classification of the property of the given object in the class identified by the given value. From this perspective, assessing the truth of a Basic Evaluation Equation is a sensible and useful task, basically free of philosophical preconditions: indeed, the idea that an entity – the property of an object in this case – can be objectively classified after a classification criterion has been set does not require one to accept any realist assumption about properties of objects or values of properties.

This interpretation can be then generalized, by refining our understanding of the two lengths involved

in the Basic Evaluation Equation above and their relations. Let us suppose that our knowledge of the structure of rod *a* leads us to model it as having a unique length at the scale of millimetres, so that $L[a] = 1.234$ m would be either true or false as such, but not at a finer scale. Were we for some reason instead expected to report a measured value for $L[a]$ at the scale of the tenths of millimetres, as above, we should admit that the measurand has to be defined with a non-null definitional uncertainty, as exemplified in Box 2.3, thus acknowledging that in this case $L[a]$ is not really *one* length, but *several* – plausibly an interval – of them. As a consequence, the Basic Evaluation Equation above, if true, should be actually meant as something like

$$L[a] \ni 1.2345 \text{ m}$$

i.e., 1.2345 times the metre is one of the lengths that rod *a* has, as the measurand $L[a]$ is defined. While this seems to be a peculiar conclusion, consider that values of quantities are sometimes conceived of as operating in mathematical models assuming the continuity (and the differentiability, etc.) of the involved functions, so that the Basic Evaluation Equation above is actually treated as if it were $L[a] = 1.234500000...$ m. However, macroscopic objects cannot have a length classified in a scale that is infinitely specific, i.e., by values with infinitely many significant digits. Hence, sooner or later the interpretation of Basic Evaluation Equations as memberships, instead of equalities, may become appropriate, where the measurand is then defined with a non-null definitional uncertainty with respect to a given scale and a value is true if it is one of the lengths of the subset / interval admitted by the definitional uncertainty. This is a plausible account of what is behind the VIM's statement that "there is not a single true quantity value but rather a set of true quantity values" (JCGM, 2012: 2.11, Note 1).

The previous construction, which has led us to reach a conclusion about what values of quantities are, explicitly relies on the additivity of length. In the next two sections we discuss how this conclusion generalizes to non-additive cases. We discuss here values of non-additive quantities, in particular those represented on interval scales, while reserving a discussion of the most general case of values of possibly non-quantitative properties to Sect. 6.5.2.

6.3.6 Beyond additivity: the example of temperature

Let us first discuss the case of temperature, as characterized and then measured in thermometric (e.g., Celsius and Fahrenheit) scales. Unlike length, temperature is not an additive quantity: that is, we do not know how to combine bodies by temperature so that the temperature of the body obtained by combining two bodies at the same temperature is twice the temperature of each of the combined bodies. This could be what led Campbell to conclude that "the scale of temperature of the mercury in glass Centigrade thermometer is quite as arbitrary as that of the instrument with the random marks" (1920: p. 359), so that "the international scale of temperature is as arbitrary as Mohs' scale of hardness" (p. 400). Were this correct, values of temperature, such as 23.4 °C, would be only identifiers for ordered classes of indistinguishable temperatures, as are values of Mohs' hardness, so that we could assess that the temperature identified as 23.4 °C is higher than the temperature identified as 23.3 °C, but not that the difference between the temperatures identified as 23.4 °C and 23.3 °C is the same as the difference between the temperatures identified as 22.9 °C and 22.8 °C. Our question is then: what *is* a value of temperature?

The starting point is the same as in the case of length: we assume to be able to compare bodies by their temperature so as to assess whether two given bodies have indistinguishable temperatures (in

analogy with the comparison depicted in Fig. 6.2) or whether one body has a greater temperature than the other.[23]

On this basis, a (non-arbitrary) scale of temperature (and therefore values of temperature) can be constructed through an empirical procedure, though, admittedly, not as simply as the one for length. As in the case of length, all assumptions that follow relate to empirical properties of objects, and non-idealities in the comparisons of such properties are not taken into account.

Let us consider a sequence a_i, $i = 1, 2, ...$, of amounts of gas of the same substance, where the i-th amount has the known mass $M[a_i] = m_i$ and is thermally homogeneous, at the unknown temperature $\Theta[a_i] = \theta_i$.[24] Let us suppose that any two amounts of gas a_i and a_j can be combined into a single amount $a_{i,j}$, such that $m_{i,j} = m_i + m_j$. It is assumed that $a_{i,j}$ reaches thermal homogeneity and that its temperature $\theta_{i,j}$ is only a function of θ_i, m_i, θ_j, and m_j (but of course the non-additivity of temperature is such that $\theta_{i,j} \neq \theta_i + \theta_j$). Finally, let us suppose that the temperatures of any two amounts of gas can be empirically compared by equality and by order, i.e., whether $\theta_i = \theta_j$ or $\theta_i < \theta_j$ or $\theta_j < \theta_i$. The hypothesis that temperature is an *intensive* property (see Sect. 1.2.1) can be tested through some preliminary checks:

- for any two amounts a_i and a_j, if $\theta_i = \theta_j$ then $\theta_{i,j} = \theta_i = \theta_j$, i.e., thermal homogeneity does not depend on mass;
- for any two amounts a_i and a_j, if $\theta_i < \theta_j$ then $\theta_i < \theta_{i,j} < \theta_j$, i.e., thermal composition is internal independently of mass;
- for any three amounts a_i, a_j, and a_k, if $\theta_i < \theta_j < \theta_k$ and $m_j \leq m_k$ then $\theta_{i,j} < \theta_{i,k}$, i.e., thermal composition is monotonic for monotonically increasing mass;
- for any three amounts a_i, a_j, and a_k, if $\theta_i < \theta_j < \theta_k$ and $m_j > m_k$ then all cases, $\theta_{i,j} < \theta_{i,k}$, $\theta_{i,j} = \theta_{i,k}$, and $\theta_{i,j} > \theta_{i,k}$, can happen.

The fact that these conditions hold may suggest the hypothesis that

$$\theta_{i,j} = \frac{m_i \theta_i + m_j \theta_j}{m_i + m_j}$$

i.e., temperatures compose by weighted average, where the weights are the masses of the composing amounts of gas. For testing this hypothesis, let us assume that three amounts of gas, a_i, a_j, and a_k, are given such that their masses m_i, m_j, and m_k are known and can be freely changed, and that $\theta_i < \theta_j$ and $\theta_i < \theta_k$. Let us now suppose that a_j and a_k, are independently composed with a_i, and m_i, m_j, and m_k are chosen to obtain that $\theta_{i,j} = \theta_{i,k}$, and therefore, under the hypothesis that temperatures compose by weighted average, that

$$\frac{m_i \theta_i + m_j \theta_j}{m_i + m_j} = \frac{m_i \theta_i + m_k \theta_k}{m_i + m_k}$$

What is obtained is a system with *two* degrees of freedom, in which one of the three unknown temperatures θ_i, θ_j, and θ_k is a function of the other two temperatures and of the three masses, i.e., $\theta_k = f(\theta_i, \theta_j, m_i, m_j, m_k)$. Were a value arbitrarily assigned to identify θ_i and θ_j (for example 0 °X and 1 °X for

[23] For the sake of simplicity, we assume that this construction is done in a context in which sufficiently clear ideas are available about what temperature is and therefore in particular how temperature and heat are related but different properties (note that sometimes temperature is considered to be the *intensity* of heat, and this justifies its non-additivity). The actual historical development of these ideas was convoluted, and some sorts of "candidate measurements" were instrumental to the clarification (see Chang, 2004, and Sherry, 2011).

[24] In the given conditions, the target here is to build a scale of temperature from a scale of mass, which is supposed to be known: this is why for any amount of gas a_i the value m_i of the mass $M[a_i]$ is considered known, whereas the value of the temperature $\Theta[a_i]$ is still undefined, and all we can say of $\Theta[a_i]$ is that it is a given temperature θ_i (where, as explained in Table 2.3, a Roman lowercase symbol, like θ_i, stands for a generic individual property and an italic lowercase symbol, like m_i, stands for a value of a property.

an X scale with values in degrees X), a value for θ_k could be computed. By choosing the two temperatures θ_i and θ_j and setting the corresponding values θ_i and θ_j, and repeating the same process with different masses m_i, m_j, m_k and a different temperature θ_k, other values of the X scale would be obtained, and the hypothesis of weighted average validated.

However, as discussed in Sect. 1.2.1, historically a key step forward was the discovery of the thermal expansion, i.e., that some bodies change their volume when their temperature changes. In metrological terms, such bodies can be exploited as transducers of temperature (see Sect. 2.3). Making a long story short, the refined treatment of these bodies – in devices that we would consider today (uncalibrated) thermometers – corroborated the empirical hypotheses that, within given ranges of volumes of given bodies,

- for a sufficiently large set $\{a_i\}$ of bodies the temperature $\Theta[a_i]$ of each body in the set and its volume $V[a_i]$ are causally connected, as modeled by a function f, $V[a_i] = f(\Theta[a_i])$,
- such that changes in temperature of each body in the set produce changes in its volume,
- and that, for each body in the set, differences in volume correspond to differences in temperature in such a way that equal differences of volume are produced by equal differences of temperature, i.e., if $V = f(\Theta)$ and $v_1 - v_2 = v_3 - v_4$ then it is because $\theta_1 - \theta_2 = \theta_3 - \theta_4$.[25]

While this development so far involves only abstract individual properties – temperatures and volumes[26] – possibly identified as properties of objects, on this basis the construction of a scale of temperatures, and therefore the introduction of values of temperature, is a relatively trivial task. According to the traditional procedure,

- two distinct temperatures are identified, θ_1 and θ_2, each of them being the common, constant temperature of a class of objects, $\theta_1 = \Theta[R_1]$ and $\theta_2 = \Theta[R_2]$, in analogy with what discussed in Sect. 6.3.4 about speed of light; θ_1 and θ_2 could be the temperatures of the freezing point of water and the boiling point of water in appropriate conditions, respectively;
- the scale built from θ_1 and θ_2 is given a name, say °C, and a number in the scale is conventionally assigned to θ_1 and θ_2, thus identifying them with values, for example 0 °C := θ_1 and 100 °C := θ_2;
- according to the hypothesis that equal differences of volume are produced by equal differences of temperature appropriate numbers in the scale are assigned to all other temperatures: for example, if $f(\theta_3) = [f(\theta_1) + f(\theta_2)] / 2$, then [0 °C + 100 °C] / 2 = 50 °C := θ_3.

The conclusion is then that values of temperature are individual temperatures identified as elements in such a scale.

6.3.7 Beyond additivity: the example of reading comprehension ability

Let us now discuss the case of reading comprehension ability (RCA), as characterized and then measured by reading tests. Like temperature and unlike length, RCA is not an additive quantity: that is, we do not know how to combine readers by RCA so that the RCA of a hypothetical "synthetic reader" is the sum of the RCAs of each of the combined readers. As above, our question is then: what

[25] The condition that this construction applies to multiple bodies / thermometers avoids the problems of radical operationalism, which would *define* temperature as what is measured by a given instrument.

[26] Differences of volumes of the relevant bodies have been assumed to be somehow observable. However, instead of operating on empirical properties it might be more convenient to measure volumes and then to operate mathematically on the measured values. Note furthermore that this role of volume as a transduced property that is a function of temperature played an important historical role, as the scientific principle at the basis of the construction of the first thermometers, but is by no means unique. An analogous presentation could be made, for example, with voltage in place of volume in reference to the thermoelectric effect.

is a value of RCA such as, say 150 RCA units? The starting point is the same as in the case of length and temperature: we assume that we can compare readers by their RCA so as to assess whether two given readers have indistinguishable RCAs (in analogy with the comparison depicted in Fig. 6.2). For example, the two readers could be asked to discuss the contents of a text passage with a human judge, and the judge could then rate the reader's relative RCAs. Now, unaided human judges may not have sufficient resolution to discriminate RCA beyond rough ordinal classes (e.g., very little comprehension, text comprehension, literal comprehension, inferential comprehension, etc.), so that one could, subject to the assumption that one used the same human judge, consider that RCA is at most an ordinal property. Apart from concerns that this may be assuming a weaker scale of RCA than possible, there are clearly serious issues of subjectivity at play in this situation: did the judge ask the same questions of the two readers, did the judge rate the responses to the questions "fairly", and would a different human judge be consistent with this one?

A key step forward was the implementation of standardized reading tests (Kelly, 1916; see Sects. 1.2.2 and 3.3.1), where readers would (i) read a text passage and (ii) answer a fixed set of questions (called in this context "items"[27]) about the contents of the passage; and then (iii) their answers would be judged as correct or incorrect, and (iv) readers would be given sum-scores (e.g., the total number of items that they answered correctly) on the reading comprehension test. Here readers who had the same sum-score would be indistinguishable with respect to their RCAs as measured by that test. Again, this would result in an ordinal scale (i.e., the readers who scored 0, the readers who scored 1, ..., the readers who scored K, for a test composed of K items), though, depending on the number of items in the set, there would be a finer grainsize than in the previous paragraph (i.e., as many levels as there are different sum-scores). This approach does address some of the subjectivity issues raised by the previous approach: the same questions are asked of each reader, and, with a suitable standardized mode of item response scoring, the variations due to different human judges can be reduced, if not eliminated altogether. However, what is not directly addressed are the issues of (a) the selection of text passages, and (b) the selection of questions about those passages. Suppose, however, that one was prepared to overlook these last two issues: one might convince oneself that the specific text passages and questions included in the test were acceptably suitable for all the applications that were envisaged for the reading comprehension test. In that case, one could adopt a norm-referenced approach to developing a scale (see Sect. 6.3.4), where the cumulative percentages of readers from a sample from a given reference-population (say, Grade 6 readers from X state in the year 20YZ) was used to establish a mapping from the RCA scores on the test to percentiles of the sample. This makes possible so-called equipercentile equating to the (similarly calculated) results of other reading comprehension tests.

Thus, at this point in the account, the conclusion is then that values of RCA are individual abilities identified as elements in an ordinal scale. It is interesting to note that the sum-scores which are the indexes used for the ranks can be also thought of as frequencies: a sum-score s, out of K total number of items, is a frequency of s in terms of the number of correct items. It can also be seen as s/K, a relative frequency (proportion) compared to the total number of items K, and, of course, relative frequencies are often interpreted as probabilities (though this move is not philosophically innocent; see, e.g., Holland, 1990, and Borsboom, Mellenbergh, & van Heerden, 2003). That is, given this set of K items, what is the average probability that the reader will get the items correct, based on the proportion that they did get correct?

To see how this makes sense, one must backtrack to our conception of RCA, as follows.

[27] As we discuss in Chap. 7, each question of a test operates as a transducer, in this case transforming the RCA of a reader to a score.

- We label as *RC-micro* an event (Wilson et al., 2019) involving a reader's moment-by-moment understanding of a piece of text. This is related to Kintch's concept of the *textbase* in his Construction/Integration (CI) model (Kintch, 2004), and refers to all component skills such as decoding (also known as word recognition), and these typically are driven from a finer to a coarser lexical granularity, i.e., a reader builds meaning from text, starting from small units (letters, sounds, words, etc.) and moving to progressively larger units. Most competent readers are not even conscious of the events at the lowest levels of granularity, unless, of course, the reader comes across a word that she does not recognize, and may have to go back to sounding it out letter by letter (i.e., grapheme by grapheme). Thus, each of these reading comprehension events can be thought of as a micro-level event that is also composed of a cascade of other more basic micro-level events, and is also contained in other micro-level events.
- In contrast, we label as *RC-macro* the events which integrate all the micro-level events which occur for a reader in the process of reading the text passage, and may integrate other conceptions beyond those, including thoughts about other texts and other ideas. This is related to the *situation* aspect of Kintch's CI model, which is integrated with the textbase, to form a deeper understanding of the text, and that is what will be stored in long term memory. Here, we might compare this to temperature, where the micro-events can be seen as the motion of individual molecules, which will each have properties of speed and direction in three-dimensional space (i.e., these would be seen as constituents of kinetic energy, a quantity different from temperature), which we cannot directly observe, and in which we are usually not interested. In contrast, the macro-level property of temperature is the integration over all of these molecular motions inside a certain body, which is what we are indeed interested in.
- This leads us to reading comprehension ability, which is the overall disposition of an individual reader to comprehend texts.
- Then, when test developers construct a RCA test, they (a) sample text passages from a body of texts, (b) design an item-universe (i.e., the population of items:[28] Guttman, 1944) of questions (items) that challenge a reader's RCA concerning parts of the text (including, of course, possibly whole texts, and across different texts), take a sample from that universe, and (c) establish rules for deciding whether the answers to the sampled questions are correct or not, resulting in a vector of judgments of responses. This then is the transduction, from RCA to a vector of scored responses to the items.
- In the tradition of *classical test theory*, as described above, the items are viewed as being "interchangeable" in the sense of being randomly sampled from the item-universe, and hence the information in the vector can be summarized as the score *s*, and, equivalently, as the relative frequency *s/K* that the reader will (on average) get an item correct.
- Alternatively, the indication could be seen as the vector of responses, thus preserving the information about individual items (such as their difficulty), and thus modeling the probability of the reader getting each of the items correct, and this is the direction followed below.

In addition, generalizability must be considered: basing the measurement of RCA on a specific test is too limiting for practical application. This was recognized by Louis Thurstone (1928), a historically important figure in psychological and educational measurement (1928: p. 547):

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently

[28] The definition (or content) of this item-universe is often referred to as the "domain".

> because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement.

To contextualize this, suppose that the RCA of readers is to be assessed using a set of items designed for reading comprehension which can be scored, as above, only as correct or incorrect. Thus, we must ask what is required so that the comparison of two readers m and n (in terms of their RCA) will be independent of the difficulty of the items that are used to elicit evidence of their relative RCAs?

Furthermore, assume that the test is composed of a set $\boldsymbol{I}$ of items. Now, two readers m and n can be observed to differ only when they answer an item differently. For any such pair of readers, m and n, there will a set of items for which they are both correct, call it $\boldsymbol{I}_c$, and a set for which they are both incorrect, $\boldsymbol{I}_i$. Then the set of items on which they differ will be $\boldsymbol{I}_d$, which is $\boldsymbol{I}$ with $\boldsymbol{I}_c$ and $\boldsymbol{I}_i$ removed – and suppose that the number of items in $\boldsymbol{I}_d$ is D. Suppose further that the number of items that reader m gets correct in the reduced set $\boldsymbol{I}_d$ is s_m, and define s_n similarly. Then, $s_m + s_n = D$, and s_m/D is the relative frequency of m answering an item correctly and n simultaneously answering it incorrectly. Thus the RCAs of m and n (in terms of the success rates of m and n) can be compared by comparing s_m/D with s_n/D, where $(s_m/D)/(s_n/D)$ is the observed proportion of reader n answering an item incorrectly and simultaneously answering it correctly. By interpreting relative frequencies as probabilities, these are then $P(m$=correct, n=incorrect) and $P(m$=incorrect, n=correct), and they can be compared using their ratio

$$\frac{P(m=correct, n=incorrect)}{P(m=incorrect, n=correct)}$$

Now, suppose that P_{mi} is the probability that person m responds correctly to item i (and equivalently for person n), so that this expression can be written somewhat more compactly

$$\frac{P_{mi}(1-P_{ni})}{(1-P_{mi})P_{ni}}$$

with the assumption of local independence, and the observation that, where there are only two responses, then the sum of the two possibilities must be 1.0.

Returning now to Thurstone's admonition, this can be translated in this context to the requirement that the equation

$$\frac{P_{mi}(1-P_{ni})}{(1-P_{mi})P_{ni}}=\frac{P_{mj}(1-P_{nj})}{(1-P_{mj})P_{nj}} \tag{1}$$

should hold for any choice of items i and j. It would be a matter of just several lines of algebra to show that

$$P_{ni}=\frac{\exp(\theta_n-\delta_i)}{1+\exp(\theta_n-\delta_i)} \tag{2}$$

where θ_n is reader n's RCA, and δ_i is item i's reading difficulty. In fact, with the probability function in eq. (2), *both* expressions in eq. (1) reduce to $\exp(\theta_m - \theta_n)$, that is, the item difficulties, δ_i and δ_j, are no longer present, which confirms that the comparison does not depend on the specific items used for the comparison, as Thurstone demanded. Note that the RCAs and item difficulties are on an interval scale (by construction). Of course, in order for the item difficulties to be eliminated from the equation, the item difficulties and the RCAs must conform to this probability model in the sense of statistical fit,

and hence this is an empirical matter that must be examined for each instrument and human population. The surprising finding about this function, in eq. (2), is that, under quite mild conditions, it is the only such function involving these parameters; this result is due to Georg Rasch, hence the function is called the "Rasch" model (Rasch, 1960/80).

The actual numbers obtained for θ_n and δ_i are termed "logits" (i.e., log of the odds, or log-odds units),[29] and are typically used to generate property values in ways similar to the way that is done for temperature units: the logits are on an interval scale, so what is needed are two fixed and socially-accessible points. One standard way is to assign two relatively extreme values: for example, one might decide that for a given population of readers, say State X for year 20YZ, the 100.0 point level might be the mean of the logits for readers in Grade 1, while a higher value, say 500.0, would be chosen as the mean for students in Grade 12: this would be suitable, for example, for a reading test used in a longitudinal context (for an expanded discussion, see, e.g., Briggs, 2019).

A second, but similar way, perhaps more suitable for test applications focused on particular grades, would be to allocate a value, say 500 as the mean for readers in Grade 6, say, and 100 as the standard deviation for the same students. Some applications also use the raw logits from their analyses – this effectively embeds the interpretation of the units in a given sample, which may be acceptable in some research and development situations, but would be difficult to justify in a broadly-used application. There is also a more traditional set of practices that (a) use an ordinal approach to classifying readers into a sequence of reading performance categories, and (b) a "norm-referenced" approach that carries out similar techniques to those just described, but using the raw scores from the reading tests rather than estimates from a psychometric model.

The conclusion is then that values of RCA are individual abilities identified as elements in a log-odds (interval) scale based on ratios of probabilities (see Freund, 2019, for a discussion of these types of scales).

6.4 The epistemic role of Basic Evaluation Equations

The conclusion reached in the previous section has an important implication for an ontology of quantities, and properties more generally as developed further in the following section. A Basic Evaluation Equation such as $Q[a] \approx x\ \mathrm{q}_{\mathrm{ref}}$ reports not just an attribution or a representation, but the claim of an indistinguishability, and in the form $Q[a] = x\ \mathrm{q}_{\mathrm{ref}}$ the claim of an equality, of individual quantities: if it is true, it informs us that two individual quantities that were known according to different criteria – as the property of an object and the multiple of a unit respectively – are in fact one and the same. In detail:

- before the relation is evaluated, we are able to identify an individual quantity q as a quantity of an object a, $Q[a]$, and a set of individual quantities q_x, each as a value $x\ \mathrm{q}_{\mathrm{ref}}$, for a given $\mathrm{q}_{\mathrm{ref}}$ and a number x varying in a given set; q and each q_x are quantities of the same kind;
- as a result of the evaluation, a number x is found such that the hypothesis is supported that the individual quantity q and the individual quantity q_x, that were known according to different criteria, are in fact one and the same, i.e., that $Q[a]$ and $x\ \mathrm{q}_{\mathrm{ref}}$ are identifiers of the same individual quantity.

As a consequence, Basic Evaluation Equations are *ontologically* irrelevant: if they are true, they simply instantiate the tautology that an individual quantity is equal to itself.[30] But, of course, their

[29] Equation 6.2 has no closed-form solution for θ and δ, hence we are not providing equations for them.

[30] This contrasts with any position which attributes a special ontic role to values. For example, Hasok Chang, as an illustration of "ontological principles ... that are regarded as essential features of reality in the relevant epistemic

widespread use is justified by their *epistemic* significance: if they are true, they inform us that two individual quantities that were known according to different criteria are in fact one and the same.

This gives us the tool to interpret a common way of presenting measurement: "the input to the measurement system is the true value of the variable [and] the system output is the measured value of the variable [, so that] in an ideal measurement system, the measured value would be equal to the true value" (Bentley, 2005: p. 3), with the consequence that "in a deterministic ideal case [it] results in an identity function" (Rossi, 2006: p. 40). Let us concede that in the deterministic ideal case the Basic Evaluation Equation is not just an indistinguishability but an equality, $Q[a] = x\ q_{ref}$. Nevertheless, the position exemplified in these quotes confuses the ontological and epistemic layers: for those who already know that $Q[a]$ and $x\ q_{ref}$ are the same quantity, the relation is an ontological identity, as is *the evening star = the morning star* for current astronomers (see Sect. 5.3.2). And in fact those who already know that $Q[a]$ and $x\ q_{ref}$ are the same quantity would have no reasons for measuring $Q[a]$. But measurement is aimed at acquiring information on a measurand, not at identically transferring values of quantities through measuring chains as is the implicit supposition behind the quotations above.

Indeed, the idea of deterministic ideal measurement as an identity function becomes understandable if it is applied not to measurement, but to transmission systems, as mentioned in Sect. 4.2.1. It is correct indeed to model a transmission system in such a way that the input to the transmission system is the value of the variable and the system output is the transmitted value of the variable, so that "in an ideal transmission system, the transmitted value would be equal to the input value" (by paraphrasing Rossi's quote above). If in fact values were empirical features of phenomena, measuring instruments could be interpreted as special transmission channels, aimed at transferring values in such a way that the transmission is performed without errors and therefore as an identity function. But a basic difference between transmission and measurement is manifest: the input to a transmission system is a value, explicitly provided by an agent who or which operates on purpose by encoding the value into the quantity of an object and sending the quantity through a channel, the purpose of which is to faithfully transfer this input. In this case, the value transmitted along the channel by the agent via the encoded quantity is in principle perfectly knowable. No such agent exists in the case of measurement, which requires a radically different description, in which values of quantities are the output, and not the input, of the process.[31]

6.5 Generalizing the framework to non-quantitative properties

The ontological and epistemological analysis proposed so far has been focused on quantities, although, as we have exemplified, much can also be done with non-additive quantities. In consistency with the VIM, we have assumed that quantities are specific kinds of properties (JCGM, 2012: 1.1), and therefore we need to work on the relation between quantities and properties in order to explore whether and how the ontology and epistemology introduced so far can be applied to properties in general. Concretely, the issue is whether Basic Evaluation Equations can involve non-quantitative properties, and if so, what are the key differences between quantitative and non-quantitative Basic Evaluation Equations.

community" and in defense of what he calls "the pursuit of ontological plausibility", mentions the "Principle of single value (or, single-valuedness): a real physical property can have no more than one definite value in a given situation." (Chang, 2001: p. 11, also presented in Chang, 2004: p. 90).

[31] This highlights the ambiguity of calling a mathematical relation among all quantities known to be involved in a measurement a "model of measurement", as the VIM definition says (JCGM, 2012: 2.48). We argue against this in Sect. 7.2.

According to a standard view in philosophy of science, developed in particular within the neopositivist tradition by Rudolf Carnap (1966) and Carl Gustav Hempel (1952), "the concepts of science, as well as those of everyday life, may be conveniently divided into three main groups: classificatory, comparative, and quantitative." (Carnap, 1966: p. 51). The VIM at least implicitly assumed this classification and adapted it to properties, defined to be either quantities or nominal properties, where the former are defined to be either quantities with unit (peculiarly, <quantity with unit> is not explicitly defined, nor given a term) or ordinal quantities. Hence according to the VIM the basic distinction is between being quantitative and non-quantitative (Dybkaer, 2013), where the demarcation criterion is <to have magnitude>: quantities are properties that have magnitude (including ordinal quantities, then), and nominal properties are properties that have no magnitude. This concept system is depicted in Fig. 6.8.

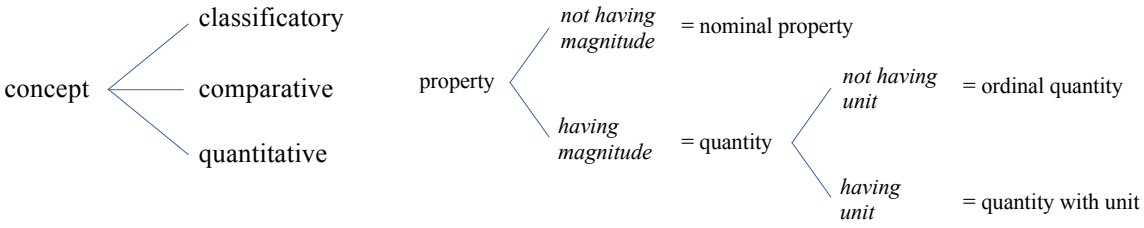


Fig. 6.8 A traditional classification of concepts (left), and its implementation in the VIM (right)

The VIM does not define what a magnitude is, but a plausible hypothesis is that "magnitude" can be generically interpreted there as "amount" or "size", so that for example the height of human beings is a quantity because it is a property that we have in amounts. This stands in contrast with properties such as blood type, which is only classificatory because we do not have "amounts of blood type". Accordingly, the phrase "the magnitude of the height of a given human being" refers to "the amount of height" of that human being: it is then, plausibly, an individual length (for a more detailed analysis of the elusive concept <to have magnitude>, see Giordani & Mari, 2012).

The simplicity of the VIM's account is attractive, but the distinction between quantitative and non-quantitative properties deserves some more analysis here, also due to its traditional connection with the discussion about the conditions of measurability, as mentioned in Sect. 3.4.2.[32]

As discussed in Chap. 4, several interpretations of measurement have been proposed, but a long tradition rooted on Euclid has connected the measurability of a property with its being quantitative. The VIM keeps with this tradition in stating that "measurement does not apply to nominal properties" (JCGM, 2012: 2.1 Note 1). As also discussed in Sect. 1.1.1 and at more length in Sect. 4.2.3, tensions related to this issue helped motivate the formation of a committee of physicists and psychologists appointed the *British Association for the Advancement of Science* and charged with evaluating the possibility of providing quantitative estimates of sensory events (see the discussion in Rossi, 2007; a more detailed analysis is in Michell, 1999: ch. 6). We might envision two distinct but complementary paths toward resolution of these tensions:

[32] This subject has been widely debated at least since the seminal analysis by Helmholtz (1887), who opened his paper by claiming that "counting and measuring are the bases of the most fruitful, most certain, and most exact of all known scientific methods". Such prestige and epistemic authority makes measurement a yearned-for target. From another perspective, "Measurement is such an elegant concept that even with [properties] apparently lacking multiples, if the [property] is capable of increase or decrease (like temperature is, for example), the temptation to think of it as quantitative is strong." (Michell, 2005: p. 289). See Chap. 7 for more on this.

- one is about the possibility of providing meaningful quantitative information for properties which are not directly or indirectly evaluated (or evaluable) by means of additive operations;[33]
- the other is about the appropriateness of broadening the scope of measurement so as to include the evaluation of non-quantitative properties.

From the beginning, both of these paths have been biased by the prestige of measurement, as witnessed by the key role attributed by some to Otto Hölder's (1901) paper, a mathematical work whose title has been translated in English as "The axioms of quantity and the theory of measurement", and about which Michell asserted that "we now know precisely why some attributes are measurable and some not: what makes the difference is possession of quantitative structure" (p. 59). In the same vein Jan De Boer claimed that "Helmholtz and the mathematician Hölder are usually seen as the initiators of the axiomatic treatment of what is often called the theory of measurement" (1995: p. 407). But even just a glance at the scope of Hölder's axioms shows that they do not relate to any experimental process, as would be expected from Helmholtz's own words – "the most fruitful, most certain, and most exact of all known scientific methods" (1887: p. 1) – and confirmed by the way the VIM defines <measurement>: a "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" (JCGM, 2012: 2.1). Indeed, Hölder himself admitted that "by 'axioms of arithmetic' has been meant what I prefer to call 'axioms of quantity'" (p. 237), so that measurement is involved in them only insofar "the theory of the measurement" is equated to "the modern theory of proportion" (p. 241), thus confirming the purely mathematical nature of the treatment.

In Sect. 3.4.2 we argued that this superposition of conditions of being measurable and being quantitative derives from a confusion between <measurement>, an empirical concept, and <measure>, a mathematical concept. The conclusion is simple to state: what is to be found in Euclid's Elements and what Hölder considered "the modern theory of proportion" is not a theory *of measurement* but a theory *of measure*, where measures are taken to be continuous quantities, where then, despite their lexical similarity, <measurement> and <measure> need to be maintained as distinct concepts (Bunge, 1973). From this, of course one might *assume* that only properties modeled as measures are measurable, but this is an assumption, not a (logical, epistemological, or ontological) necessity: it is what we take to be the position of what Michell (1990) calls "the classical theory of measurement", as rooted in Euclid's geometry, but his sharp tenet that "without ratios of magnitudes there is no measurement" (p. 16) cannot be maintained without this strong and basically arbitrary assumption.

The problems generated by this confusion are not just lexical or semantic. A well grounded distinction between quantitative and non-quantitative properties would be a key target, at least as a means to identify and justify possible differences in inferential processes and their results, and therefore in the kind of the information they produce. The basic intuition about the distinction remains, e.g., that individuals can be compared in such a way that the height of a person can be one third greater than the height of another, or a difference on an interval scale can be one third greater than another difference, whereas the blood type of that person cannot be one third greater of the blood type of another. This intuition needs a persuasive explanation, which ultimately would be beneficial for a better identification of the conditions of measurability. In fact, the mentioned confusion is a good reason for developing this analysis as a key component of an ontology and an epistemology *of properties*: once an appropriate classification of types of properties has been established, whether only

[33] We already followed this path in the discussion about values of temperature and of reading comprehension ability in Sects. 6.3.6 and 6.3.7.

quantities are measurable might be thought of as simply an arbitrary lexical choice (Mari et al., 2017).[34]

A basic framework on this matter was proposed by Stanley Smith Stevens (1946), with his well-known classification of what he called "scale types", and since then his distinction between nominal, ordinal, interval, and ratio scales has been widely adopted (for an early, extended and clear presentation, see Siegel, 1956: ch. 3), and variously refined.[35] Such a framework was conceived as dealing with scales *of measurement*, given the perspective that "measurement, in the broadest sense, is [...] the assignment of numerals to objects or events according to rules" (p. 677), explicitly drawing from Campbell's seminal representationalist statement that "measurement is the assignment of numerals to represent properties" (Campbell, 1920: p. 267; see also the related discussion in Sect. 4.2). From the perspective of the present analysis Stevens' "broadest sense" is indeed too broad, if considered to be specifically related to measurement. Rather, what is interesting in his classification is more correctly understood by considering it as related to scales of property *evaluation*, thus disentangled from issues about measurability. We have then to deal with two interrelated issues:

- to which entities does the feature of being quantitative or non-quantitative apply?
- how should the condition of being quantitative or non-quantitative be defined?

But are the terms "nominal", "ordinal", and so forth best understood as referring to types *of properties*, or of *evaluations*? And, in consequence, how should such types be defined?

6.5.1 The scope of the quantitative/non-quantitative distinction

The first question for us to consider is about the scope of the classification between nominal, ordinal, interval, and ratio – let us call it NOIR, from the initials of the four adjectives – that is, what does it apply to, and therefore what does NOIR classify? At least two positions are possible. According to one, NOIR is about assignments of informational entities to objects: nominal, for example, is a feature of the way numerals (in Stevens' lexicon, i.e., "names for classes", 1946: p. 679) are assigned to individuals with respect to their blood type. This is how Stevens introduced it, thus considering for example blood type to be evaluated on a scale that is nominal. According to another position, NOIR is about the properties themselves: nominal, for example, is a feature of blood type. This is how, for example, the VIM uses it, thus considering blood type to be a nominal property. Hence given a Basic Evaluation Equation such as

blood type[*individual x*] = A in the ABO system

[34] Given this, the reader will not find here the proposal of a clear-cut criterion to distinguish between quantities and non-quantities. At least since Hölder's (1901) paper, several axiomatizations of quantities have been proposed (e.g., Suppes, 1951; Mundy, 1987; Suppes & Zanotti, 1992), and choosing among them is not relevant here. On this matter a general issue is whether order is sufficient for a property to be considered a quantity. While the sources just cited all answer this question in the negative, more encompassing positions are possible, such as Ellis', according to whom "a quantity is usually conceived to be a kind of property. It is thought to be a kind of property that admits of degrees, and which is therefore to be contrasted with those properties that have an all-or-none character." (1968: p. 24). By using the term "ordinal quantity", the VIM adopted the same stance (JCGM, 2012: 1.26): ordinal properties are considered to be quantities. This multiplicity is one more reason not to fall in the trap of what Abraham Kaplan called the "mystique of quantity" (1964: p. 172).

[35] For example, Nicholas Chrisman mentions the following ten "levels [which] are by no means complete", where for each level the "information required" is specified (1998: p. 236): 1: Nominal (definition of categories). 2: Graded membership (definition of categories plus degree of membership or distance from prototype). 3: Ordinal (definition of categories plus ordering). 4: Interval (definition of unit and zero). 5: Log-interval (definition of exponent to define intervals). 6: Extensive ratio (definition of unit – additive rule applies). 7: Cyclic ratio (unit and length of cycle). 8: Derived ratio (units – formula of combination). 9: Counts (definition of objects counted). 10: Absolute (type: probability, proportion, etc.).

being nominal is considered to be either[36]

- a feature of the evaluation that produces the equation, according to the first position, or
- a feature of the general property that is involved in the equation, according to the second position.

By interpreting them in a representational context, Michell presents these two positions as being about *internal* and *external* representations, respectively (Michell, 1999: p. 165–166, from which the quotations that follow are taken – for consistency, everywhere the term "attribute" used by Michell has been substituted with "property"). According to Michell, an internal representation

> occurs when the [property] represented, or its putative structure, is logically dependent upon the numerical assignments made, in the sense that had the numerical assignments not been made, then either the attribute would not exist or some component of its structure would be absent.

Thus, it is internal to the evaluation. An external representation is instead

> one in which the structure of some [property] of the objects or events is identified independently of any numerical assignments and then, subsequently, numerical assignments are made to represent that [property]'s structure

where the adjective "external" is explained by Michell as the hypothesis that the property

> exists externally to (or independently of) any numerical assignments, in the sense that even if the assignments were never made, the [property] would still be there and possess exactly the same structure.

Thus, it is external to the evaluation. In summary (Giordani & Mari, 2012: p. 446),

- an internal representation is an evaluation that *induces* a structure, whereas
- an external representation is an evaluation that *preserves* a structure.

The examples proposed by Michell are interesting, and useful for better understanding what is at stake with this distinction. He exemplified *external* representations (i.e, such that NOIR is a feature of properties) by means of hardness:

> Minerals can be ordered according to whether or not they scratch one another when rubbed together. The relation, x scratches y, between minerals, is transitive and asymmetric and these [features] can be established prior to any numerical assignments being made.

The idea is then that once a property-related criterion of comparison has been identified (in this case, mutual scratching), the outcomes of property-related comparisons do not depend on the way they are represented: the conclusion would be that hardness *is* ordinal (or, more correctly, that hardness is at least ordinal). As the example suggests, this seems to be based on the assumption that, for an external representation to be possible, properties of objects must be empirically comparable according to some given conditions, and the outcome of the comparison must be observable, as in the paradigmatic case of mass via a two pan balance. This condition was embedded in the representational theories of measurement under the assumption that the availability of an empirical relational system is a precondition of measurement.

[36] We are not concerned here with whether the general property applies to single entities or to pairs, triples, etc. of them, and therefore whether – in the traditional terminology – it is a *property* or a *relation* (see Sect. 5.2.3).

Michell proposes two examples of *internal* representations (i.e., such that NOIR is a feature of representations rather than properties). The first one is about

> an extreme case [...] of assigning different numbers to each of a class of identical things (say, white marbles) and on that basis defining a [property]. The [property] represented by such assignments would not be logically independent of them and, so, had they not been made, the [property] would not exist.

This is indeed the extreme case of an assignment claimed to be a representation but that does not represent anything, being only a means of object identification: it is not even a property evaluation, given that there is no property to evaluate, in the specific sense that a Basic Evaluation Equation cannot be written because there is not a general property to be evaluated of the considered objects.[37] We may then safely ignore this case, and consider the second, "less extreme" example,

> where an independent [property] may exist, but the structure that it is taken to have depends upon numerical assignments made. For example, people may be assigned numbers according to nationality (say, Australian, 1; French, 2; American, 3; Belgian, 4; etc.) and then the [property] of nationality may be taken to have the ordinal structure of the numbers assigned. In this case, had numerical assignments not been made, the [property] (nationality) would still exist but the supposed ordinal structure would not.

This is a case in which Stevens' framework proves to be non-trivially applicable. While it is always possible to adopt numbers for representational means, the numerical relations do not necessarily relate to empirical relations among the objects:[38] in this case, although it is representable by means of ordered entities, nationality is not itself ordinal.[39] These two examples show why we do not see the category of internal representations as relevant to measurement.

Hence, in our view, *the evaluated property exists and has features that are independent of its possible representations*: an evaluation is expected to preserve the structure of the property, not to induce a structure on the property.[40]

Given the controversial nature of Stevens' framework, it may be worth noting that this has nothing to do with setting constraints on ways of representation and of related data processing, such as, say, proscribing against computing the mean value of a set of numbers that encode nationalities. Along the same lines as Lord (1953), Velleman and Wilkinson (1993) emphasized the importance of not imposing such constraints, given that "experience has shown in a wide range of situations that the application of proscribed statistics to data can yield results that are scientifically meaningful, useful in making decisions, and valuable as a basis for further research" (p. 68) ("proscribed statistics" are those statistics that are not "permissible" in the vocabulary of Stevens, 1946: p. 678). In fact, measurement science does include some "proscriptions", such as the condition of dimensional analysis that only

[37] This is analogous to the unfortunate example given by Stevens about "the numbering of football players for the identification of the individuals" (1946: p. 678): identification is not property evaluation, and so the mocking critique by Lord (1953) rightly applies to this example.

[38] This is why in the definition of <quantity> given by the VIM – "property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference" (JCGM, 2012: 1.1) – the last part "that can be expressed as a number and a reference" is not an actual specification, and could be removed.

[39] Of course, nationality-related orderings can be defined; for example, names of nations are ordered alphabetically and nations are ordered by the size of their population, but the former is a relation among linguistic terms and the latter is a relation among cardinalities of sets: neither of them involves the property of nationality, in the sense that one's nationality is not a linguistic entity, nor is it a number.

[40] This is thus in sharp contrast with radically constructionist presentations, such as Kaplan's view that "the order of a set of objects is something which we impose on them. We take them in a certain order; the order is not given by or found in the objects themselves." (1964: p. 180).

values of quantities of the same kind can be added. Nevertheless, the idea that through data analysis something can be discovered also about the structure of the evaluated properties is not problematic *per se*. The point is that if the property under consideration is evaluated based on (for example) purely ordinal comparisons (as in the case of hardness), the values that are obtained cannot be expected to convey more than ordinal information, exactly as prescribed by Stevens' framework and its refinements (an example of which is mentioned in Footnote 32). In this view, what Stevens introduced as the set of "permissible" functions is better understood as a matter of algebraic invariance and meaningfulness under scale transformation (Narens, 2002), and therefore of uniqueness of the scale itself.

A summary can be presented simply as follows:

- the *representation* of properties of objects, or the representation of objects as such, is an unconstrained process, and anything could in principle be used as a means of representation;
- the *evaluation* of properties of objects is a process that is expected to produce values of the evaluated properties;
- the *measurement* of properties of objects is a specific kind of evaluation.

From this point of view, we consider the emphasis on representation that has usually accompanied NOIR to be misleading: the position that assignments are representations that do not represent anything (Michell's "internal representations") is void, and the interesting question is instead whether NOIR is about

- ways of evaluating properties, or
- properties as such,

where in both cases the claim is that *there is* a property under consideration, having structural features which do not depend on whether or how it is represented. While Stevens, who was inclined toward operationalism, was candid about this alternative – "the type of scale achieved depends upon the character of the basic empirical operations performed" (Stevens, 1946: p. 677) – and consistently considered NOIR a feature of scales, we still have to further explore regarding this subject.

6.5.2 From values of quantities to values of properties

We have assumed so far that the concept <evaluation> applies not only to quantities but also, and more generally, to properties. This has been the justification for adopting the same structure for the Basic Evaluation Equation for both quantitative cases, e.g.,

length[*rod a*] = 1.2345 m

and

reading comprehension ability[*individual b*] = 1.23 logits (on a specific RCA scale)

and non-quantitative cases, e.g.,

blood type[*individual c*] = A in the ABO system

Hence in the generic structure

property of a given object = value of a property

A in the ABO system is an example of a value of a property, just as 1.2345 m is an example of a value of a quantity. While it is acknowledged (for example by the VIM) that quantities are specific types of properties, whether values of quantities can be generalized and thus applied to non-quantitative properties is a much less considered subject, as is the related issue of what a value of a property is. For example, the VIM defines <value of a quantity> (JCGM, 2012: 1.19), but does not define <value of a

property> even though it deals with non-quantitative properties, termed "nominal properties" (JCGM, 2012: 1.30). Hence the problem for us to consider here is whether the Basic Evaluation Equation is meaningful only in the specific case of quantities, i.e., in the case

quantity of a given object = value of a quantity

or is also able to convey knowledge about non-quantitative properties.

Our construction of values of quantitative properties, presented in Sect. 6.3, relies on their empirical additivity, in the example of length, or on the invariance of their empirical difference, in the examples of temperature and reading comprehension ability, conditions which do not hold for non-quantitative properties. Our concept of shape, for example, is indeed such that the ideas of "adding objects by their shape" or "subtracting objects by their shape" are meaningless, in the sense that the shape of an object obtained by somehow composing two other objects is in general not additively related to the shapes of the composing objects (in fact, shapes are not even ordered: for example, it is meaningless to say that a cube is more or less than a cylinder). Hence, the interpretation of the Basic Evaluation Equation for quantities, according to the Q-notation, such that $Q[a] / [Q]$ is a number (see Sect. 5.3.4) does not apply to non-quantitative properties: there are no "shape units" in this case, nor can shapes be compared by their ratio.

At a more fundamental level, however, the idea of conveying information on properties like blood types of individuals or shapes of objects by means of "values of blood type" and "values of shape" maintains its meaning and relevance, as when we say that a given rod is cubic or cylindrical, phraseological means for "whose shape is cube" and "whose shape is cylinder", which immediately leads to a formalization such as *shape*[*rod a*] = cube and *shape*[*rod a*] = cylinder. Given their analogy in structure with *length*[*rod a*] = 1.2345 m, where 1.2345 m is a value of length, the conclusion seems to be that cube, cylinder, and so forth may be considered to be values of shape.

However, this is not completely correct, given an important difference between the two cases. Indeed, the value 1.2345 m includes, via the unit metre and the reported number of significant digits, information on the set of possible values from which 1.2345 m has been chosen, i.e., the set is the non-negative multiples of the metre: it is 1.2345 m and not 1.2346 m, and so on; it is 1.2345 m but we are unable to distinguish between 1.2345 m and 1.23451 m, and so on. Choosing a unit and a (finite) number of significant digits corresponds to *introducing a classification* on the set of the lengths, in which each value identifies one class. Hence, selecting a value of length conveys both the information that (i) the class of lengths identified by that value has been selected, and (ii) all other classes (identified by all other multiples of the unit) have not been selected. In a relation such as *shape*[*rod a*] = cube this second component is missing.[41]

In order to improve the structural analogy between *length*[*rod a*] = 1.2345 m and *shape*[*rod a*] = cube, the set of the possible shapes, of which cube is one element, needs to be declared as part of the equation: it might be, e.g., {cube, any other shape} or {cube, cylinder, cone, sphere, any other shape}, thus showing that the report that rod *a* is cubic conveys different information in the two cases. We may call the set of the possible shapes a *reference set*, R, so that an example of the Basic Evaluation Equation in the case of a nominal property such as shape is[42]

shape[*rod a*] = cube in R

[41] This lack of a context – seen, for example, in that reporting that the shape of a given object is cube does not in itself provide any hint about what other shapes the object might have had – is a problem in particular for computing the quantity of information obtained by a value. According to Claude Shannon (1948), this is related to the probability of selecting that value, which in turn supposes knowledge of the underlying probability distribution. We further discuss this fundamental idea by Shannon in Sect. 8.1, in terms of quantity of (syntactic) information conveyed by measurement.

[42] This is clearly analogous to the way information is reported in ordinal cases, such as Mohs' hardness, e.g., *hardness*(*given sample*) = 5 on the Mohs scale.

where cube in R is then the example of a value of a property. Indeed, the same structure may also be used for quantities, e.g.,

length[*rod a*] = 1.2345 in metres

i.e., the value is 1.2345 in the classification of lengths generated by the metre and its multiples, but in this case additivity of length permits the more informative interpretation that the class identified as 1.2345 in that classification corresponds to a length which is 1.2345 times the metre.

Hence the concept <value> is not bound to quantities: non-quantitative properties also have values, and any such value is an individual property identified as an element of a given classification of comparable individual properties,[43] such that if the classification changes, and therefore a different reference set is used, another value may be obtained for the same property under evaluation. Under these conditions the previous considerations about values of quantities can be correctly generalized to values of properties: first, choosing a set of values for blood type or shape corresponds to introducing a classification on the set of the blood types or the shapes, in which each value identifies one class, and, second, Basic Evaluation Equations also apply to non-quantitative properties and, if true, they convey much richer information than just representation: they state that the property of an object and the value of a property are the same individual property.

Box 6.2 – Evaluation scales

In this context the question of *what is a scale* – we prefer to refer here to the more generic concept <evaluation scale> than to <measurement scale> given that what follows is not only and specifically about measurement – can be straightforwardly discussed.

Let us consider the concrete case of Mohs' scale of mineral hardness. The observation that any pair of minerals x and y can be put in interaction so that either one scratches the surface of the other or neither scratches the surface of the other, is accounted for as a relation between their hardnesses H, $H[x] > H[y]$, or $H[x] \approx H[y]$, or $H[y] > H[x]$, and ten equivalence classes, C_1, C_2, …, C_{10}, were thus empirically identified, such that if $H[x] \in C_i$ and $H[y] \in C_j$, and $i > j$, then $H[x] > H[y]$. However, dealing with equivalence classes for conveying information about properties of objects is inconvenient: rather, each class can be uniquely identified via a one-to-one mapping f from the set of the equivalence classes to a set of identifiers (for example, the set of natural numbers), where the condition of injectivity guarantees that the information about hardness-related distinguishability is not lost in the mapping. Furthermore, since mutual scratching induces an ordering on the set of the equivalence classes C_i, the mapping f may be defined so as to maintain such structural information, i.e., if $H[x] > H[y]$ and $H[x] \in C_i$ and $H[y] \in C_j$, then $f(C_i) > f(C_j)$, where then $H[x] > H[y]$ is an empirical relation about properties (hardnesses, in this example) of objects and $f(C_i) > f(C_j)$ is an informational relation about identifiers of equivalence classes. The two conditions of injectivity and structure preservation make f an isomorphism (surjectivity is an immaterial condition here): it is a scale, i.e., *an isomorphism from equivalence classes of property-related indistinguishability to class identifiers*.

For a given a set of equivalence classes $\{C_i\}$ established for a property P, the condition that the scale f be an isomorphism constrains the set of class identifiers, i.e., the range of f, except for an isomorphism. In the example, the range of f is usually taken to be the set of the first 10 natural numbers, so that f(equivalence class of talc hardnesses) := 1, f(equivalence class of gypsum

[43] As a consequence, the possible concern that only numbers (or numerals) count as values of properties is unjustified. This also shows that the values of non-quantitative properties are not merely "symbols" or "names". The discussion in Sect. 6.2.1 about values of quantities applies more generally to values of properties.

hardnesses) := 2, and so on, but any other ordered set of 10 elements could be equally chosen, say the sequence ⟨a, b, …, j⟩, where the mapping 1 → a, 2 → b, … is usually called a *scale transformation*, though a better term is *scale identifiers transformation*, given that the empirical component of the scale is left untouched and only the scale identifiers are (isomorphically) changed.

Note that the definition of a scale is a normative statement that establishes which equivalence class is identified by which identifier. As such, it is not an equation (and therefore not a Basic Evaluation Equation), and it is neither true nor false. In the example above, it is written then

$$\forall i \in \{1, 2, \ldots, 10\}, f(C_i) := i$$

(where the notation "$x := y$" means that x is defined to be y, not that x and y are discovered to be equal). From this characterization it is simple to see why for the same general property on can construct scales that are *distinct* (in the specific sense of being non-isomorphic):

- the criterion that defines the equivalence classes C_i could be changed, so that a new set of equivalence classes implies a new mapping f; for example, that in the case of refining hardness classes this could lead to non-integer identifiers like 4.5 for steel;
- the structure that needs to be preserved by the mapping f could be changed, so that a new isomorphism has to be obtained that preserves an algebraically stronger or weaker structure; an example of the first case is the historical development of the measurement of temperature when the absolute zero was discovered, allowing measurement of temperatures in the thermodynamic (Kelvin) scale and not only in a thermometric (Celsius, Fahrenheit) scale; an example of the second case is the daily measurement of temperature whenever measured values are reported in a thermometric scale, thus forgetting the available information of the "natural", absolute zero.

From a scale f a function g can immediately be derived – "lifted", in the algebraic jargon – mapping properties of objects belonging to equivalence classes to class identifiers: if $P[x] \in C$, then $g(P[x]) := f(C)$ (for example, since f(equivalence class of talc hardness) := 1, then g(hardness of a given sample of talc) = 1, i.e., in the Mohs scale the hardness of any sample of talc is identified by 1, and so on).[44] Of course, properties of distinct objects may be indistinguishable from each other, i.e., may belong to the same equivalence class, and therefore may be associated to the same identifier via a function g. Hence, while f is one-to-one, g is many-to-one, and therefore a homomorphism, that may be called a *scale-based representation of the properties of objects*. In contest with the statement $f(C_i) := i$ considered above, a relation like $g(P[x]) = i$ is typically not about constructing a scale but, after a scale f has been constructed, about using it, with the aim of identifying the property $P[x]$ as an element of an equivalence class, the one identified by i. As such, it is an equation, that is either true or false, depending on whether actually $P[x] \in C_i$, if $f(C_i) := i$.

In summary, a scale is built under the assumption that some relations (at least indistinguishability, but possibly order and more) among comparable properties of objects are given, and is introduced to organize and present in a standard way the information about such relations. As a consequence, if these relations change, the scale could be changed in turn.

[44] The usual mathematical construction follows the opposite direction, by showing that any function g: $A \rightarrow B$ induces a partition, i.e., a set of equivalence classes, on its domain A, such that a_i, $a_j \in A$ belong to the same equivalence class if and only if $g(a_i) = g(a_j)$, where the set of equivalence classes is called the quotient set of A under the equivalence relation.

6.5.3 Property Evaluation Types

Given this broad characterization of what values of properties are, it is now clear that the four conditions that in Chap. 2 we have proposed as necessary for a process to be considered a measurement can also be fulfilled by the evaluation of a non-quantitative property: it may be a process that is empirical (Sect. 2.2.1) and designed on purpose (Sect. 2.2.2), whose input is a property of an object (Sect. 2.2.3), and that produces information in the form of values of that property (Sect. 2.2.4).

However, as previously noted, such conditions are not claimed to be also sufficient. In other words, since measurement is a property evaluation but not all property evaluations are measurement, the fact that conditions that are necessary for measurement apply to the evaluation of non-quantitative properties is still not sufficient to conclude that non-quantitative properties are measurable. While Chap. 7 is devoted to proposing our account of the structural conditions that characterize measurement, it is time now to come back to the issue of whether NOIR is about ways of evaluating properties or about properties as such.

The question of the scope of NOIR, as elaborated in Sect. 6.5.1, is in fact about the alternative between a more modest instrumentalist, epistemological position, which assumes that we can only characterize evaluations (and more specifically measurements) rather than properties as such, and a stronger realist, ontological position, according to which we can instead say something about properties themselves, plausibly also on the basis of what we learn in the process of evaluating them. Of course, the more modest position is also safer, and seems to be more consistent with falsificationism (Popper, 1959) and better able to take into account the fact that scientific revolutions (Kuhn, 1969) can annihilate bodies of knowledge that were deemed to be established: given the always revisable status of our hypotheses about the empirical world – as illustrated, for example, by the historically well-known cases of phlogiston and the caloric – wouldn't it be wiser to renounce any ontological claim about the structure of properties as such?

Let us explore the issue in the light of the assumption of two conditions of *information consistency* for an evaluation (Giordani & Mari, 2012: p. 446):

(C1) for each relation among properties of objects there is a relation among values such that the evaluation preserves all property relations: this guarantees that the information empirically acquired is maintained by the evaluation;

(C2) only relations among values that correspond to relations among properties of objects are exploited while dealing with values: this guarantees that the information conveyed by values is actually about the evaluated properties.

In summary, *values should convey all and only the information available on the evaluated properties*. This is plausible, to say the least: given that values are what an evaluation produces, they should report everything that was produced, and that otherwise would be lost (C1), but nothing more than that, to prevent unjustified inferences (C2). These two conditions deserve some more consideration.

Condition (C1) seems obvious, particularly in the context of representational theories of measurement where it may be considered the premise of representation theorems.[45] For example, if the property of an object is compared with the property of another object and the former is observed to be greater than the latter, the value of the former should be greater than the value of the latter. However, the meaning of (C1) is based on the non-trivial acknowledgment that properties of objects may also be

[45] For example, Fred Roberts describes what he calls "the representation problem" as follows: "Given a particular numerical relational system Y, find conditions on an observed relational system X (necessary and) sufficient for the existence of a homomorphism from X into Y." (1979: p. 54).

compared independently of their evaluation, and therefore that the comparison has features which are independent of the evaluation. The condition that the property of one object is greater than the property of another object might be in some sense observable, and in this case does not require such properties to be evaluated. This gives support to the position that NOIR is a feature not only of the ways in which properties are evaluated, *but of properties as such*, via what we know about the ways in which they can be compared.[46] In Michell's words, "the existence of the empirical relations numerically represented must be logically independent of the numerical assignments made. That is, these empirical relations must be such that it is always possible (in principle, at least) to demonstrate their existence without first making numerical assignments" (Michell, 1999: p. 167). For sure, any such ontic claim may be updated, and in particular improved – for example when a metric is discovered to apply to what was previously considered to be a non-quantitative property – but this is just in agreement with the general understanding that empirical knowledge is always revisable.

Condition (C2) has more complex implications: how can we be sure that a relation among values does *not* correspond to a still-unobserved relation among properties of objects? The point here is not about accepting or refusing "proscriptions", in the sense of Velleman and Wilkinson (1993) and as already discussed in Sect. 6.5.1, but about acknowledging that *through evaluation some features of properties might be discovered*. For example, historically, the idea that temperature can be evaluated on an interval scale was formulated as the result of its evaluation by means of thermometers, not via the comparison of temperatures of objects in terms of their distances / intervals. As documented by Chang (2004), a crucial problem was in the confirmation of the preliminary hypothesis that the evaluation is linear (in this case, that thermometers have a linear behavior in transducing temperatures to lengths), so that divisions in the scale of values (in this case, of length in the capillary) can be treated as evidence of correspondingly proportional divisions in the scale of properties of objects (in this case, of temperatures).[47] Such an inference is then justified on the basis of *the structure of the evaluation*, in the case of thermometers realized by the transduction effect of thermal expansion. And thus, for a property already known to be comparable in terms of order, appropriate conditions on the way the property is evaluated may help justify the hypothesis that distances / intervals, and therefore units (though without a "natural" zero) are also meaningful. Such a general characterization is not limited to physical properties: Indeed, this can be understood as the rationale of simultaneous conjoint measurement (Luce & Tukey, 1964) and Rasch measurement (Rasch, 1960), as also discussed in Sect.

[46] It seems paradoxical that representationalism – a weak position about the epistemic state of measurement, as also discussed in Chap. 4 – assumes some strong ontic requirements on properties.

[47] This hypothesis of linearity can be empirically corroborated by ascertaining that different temperatures produce proportional changes in different thermometers, operating according to different transduction effects. Four conceptual (though not necessarily historical) stages may be envisioned to such a process:

1. A property is known only via a single transduction effect: for example, temperature can be transduced to a single kind of thermometric fluid (e.g., alcohol). In this case, the hypothesis of linearity is only grounded on the meta-hypothesis of simplicity.

2. A property is known via multiple transduction effects related to the same transduction principle: for example, temperature can be transduced to different kinds of thermometric fluid (e.g., alcohol and mercury). In this case, if (for example) it were discovered that the temperature that produces the midpoint in volume between the volumes produced by two fixed points (e.g., the freezing and boiling points of water at sea level) is the same for different fluids, the hypothesis of linearity gains more plausibility. (As it happens, this is *not exactly* the case for mercury and alcohol.)

3. A property is known via multiple transduction principles: for example, temperature can also be transduced to electric tension, via the thermoelectric effect. In this case, if (for example) it were discovered that the temperature that produces the midpoint in volume between the volumes produced by the two fixed points and the temperature that produces the midpoint in tension between the tensions produced by the same fixed points is the same for different bodies, the hypothesis of linearity gains more plausibility.

4. A property becomes part of a nomic network (see Sect. 6.6.2), if for example, a law is discovered that connects proportional differences of temperature of a given body to transferred heats, the hypothesis of linearity gains even more plausibility.

4.4.1:[48] the fact that the evaluation fulfills given conditions leads one to infer that the evaluated property may have a structure richer than the observed one.

The attribution of an unobserved feature to a property is clearly an important and consequential move. While according to condition (C1) NOIR would be considered a feature of properties, known through their means of comparison, condition (C2) suggests a more cautious position, that NOIR is explicitly a feature of evaluations, and only in a derived and more hypothetical way a feature of evaluated properties. That is why we propose that NOIR are examples of *Property Evaluation Types* (Giordani & Mari, 2012). This is along the same lines as Stevens' "types of scales of measurement", but with the acknowledgment that such types are more generally features of evaluations, and not only of measurements. This position allows us to take into account the fact that the same property may be evaluated by means of evaluations of different types,[49] so that the usual property-related terms – "nominal property", "ordinal property", etc. – are meant as shorthands for something like "property that at the current state of knowledge is known to be evaluable on a nominal scale at best", and so on. Even the very distinction between quantitative and non-quantitative properties has, then, this same quality: as the historical development of the measurement of temperature shows, a property that we can evaluate only in a non-quantitative way today might tomorrow also become evaluable quantitatively.

On this basis, we may finally devote some consideration to our most fundamental problem here: the conditions of existence of general properties.

6.6 About the existence of general properties

A basic commitment at the core of our perspective on measurement is that it is both an empirical and an informational process, aimed at producing information about the world, and more specifically, about properties of objects. A direct consequence of this view is that *a property cannot be measured if it does not exist as part of the empirical world*; that is, the empirical existence of a property is a necessary, though not sufficient, condition for its measurability (Mari et al., 2018). This statement may seem so obvious as to approach banality, but it has some less obvious features and consequences worthy of further exploration. In particular, one may ask: how can we *know* that a property exists? Stated alternatively, under what conditions is a claim about the existence of a property justified? And, more specifically, what does a claim of existence of a general property assume?

This section is dedicated to an analysis of this question, beginning with some conceptual house cleaning, related to the distinction between empirical properties and mathematical variables.

6.6.1 Properties and variables

We have proposed that empirical properties are associated with modes of empirical interaction of objects with their environments. To help sharpen up this statement, let us consider the distinction between empirical properties and mathematical variables. An (existing) empirical property can, in principle, be modeled by a mathematical variable; indeed, this is one of the primary activities involved

[48] For an extensive presentation of conjoint measurement, see Michell (1990: Chapter 4), where conjoint measurement is introduced as a "general way [...] in which evidence corroborating the hypothesis [that a property is quantitative] may be obtained" (p. 67). In the light of the discussion in Sect. 3.4.2, a method of quantification is not necessarily a method of measurement: hence a more correct term for conjoint measurement would be "conjoint quantitative evaluation".

[49] For example, the (length of the) diameter of objects, whose evaluation is usually of ratio type, may be evaluated by means of a sequence of sieves of smaller and smaller opening, where each sieve is identified by an ordinal value and the evaluation sets the diameter of each object to be equal to the value of the last sieve crossed by the object. Such an evaluation is then only ordinal.

in a measurement process, as described in more detail in the following chapter.[50] However, it would be fallacious to *conflate* empirical properties and mathematical variables, or to assume that the presence of either implies the existence of the other: there can be empirical properties without corresponding mathematical models (for example, because we are unaware of the very existence of such properties; e.g., blood type prior to 1900), and there can be mathematical variables without corresponding empirical properties (for example, the variables in generic mathematical equations such as $y = mx + b$).

Although this distinction may seem obvious when presented in these terms, conventions in terminology and modes of discourse may sometimes obfuscate it, as when the term "variable" is used to refer both to an empirical property and a mathematical variable (which is common in the literature on "latent variable modeling", for example; see McGrane & Maul, 2020), or when, as described in the GUM, "for economy of notation [...] the same symbol is used for the [property] and for the random variable that represents the possible outcome of an observation of that [property]" (JCGM, 2008: 4.1.1).

As a consequence, it cannot be assumed out of hand that any given feature of a mathematical variable is shared by the empirical property that the variable claims to model. For example, some physical quantities are customarily and effectively modeled as real-valued functions – a precondition for modeling the dynamics of such quantities by means of differential equations – but assuming that all features of real numbers apply to the quantities they purport to model could, for example, lead to the conclusion that a given quantity is dense in the way that real numbers are, which in many cases is known to be false, as in the case of quantized quantities such as electrical charge. Analogously, properties are customarily and effectively modeled as continuous random variables for a variety of purposes, but, again, this does not guarantee that all features of continuous random variables hold true for the modeled properties (see also, e.g., Borsboom, 2006; McGrane & Maul, 2020), even for models that fit the data according to commonly-accepted criteria (see, e.g., Maraun, 1998, 2007; Maul, 2017; Michell, 2000, 2004).

With respect to the confusion between a knowable entity and what we know of it (i.e., the concept that we have of it), a particularly pernicious class of properties are those considered to be, in some sense, *constructed*, as was previously discussed in Sect. 4.5: one might infer from the fact that "concepts such as compassion and prejudice are [...] created from [...] the conceptions of all those who have ever used these terms" that they therefore "cannot be observed directly or indirectly, because they don't exist" (Babbie, 2013, p.167). This fallaciously conflates the *concepts* we have of psychosocial properties such as compassion with the empirical *referents* of those concepts.[51] That is, if compassion, prejudice, and other psychosocial properties have ever been measured, what was measured was a property (of an individual or a group), rather than a concept of (or term for) that property.

Thus, whether a given property is defined in purely physical terms or not, the critical question is how we *know* that a property exists, and therefore that it meets at least the most basic criterion for measurability. What, in other words, justifies one's belief in the existence of a property?

[50] The identification of the conditions that make such modeling possible is one of the primary contributions of the representational theories of measurement, the stated aim of which is "to construct numerical representations of qualitative structures" (Krantz et al., 1971: p. xviii). (Perhaps peculiarly, in the terminology of representationalism, the term "qualitative" is used to refer to the structure of properties even when they are quantities.)

[51] Again, as discussed in Sect. 4.5 (and at more length in a variety of sources such as Mislevy, 2018), there are many important differences in the ontological character of psychosocial properties compared to (classical) physical properties, including the facts that their existence depends on human consciousness (with all the ontological challenges this entails; see, e.g., Dennett, 1991; Kim, 1998; Searle, 1992), and perhaps also on social groups and the actions thereof (with all the ontological challenges that entails; see, e.g., Searle, 2010). However, the key point remains: none of these differences imply that psychosocial properties are not part of the empirical world, any less so than physical properties.

6.6.2 Justifications for the existence of properties

There are many ways in which a claim about the existence of an empirical property could be justified, but given the *empirical* nature of the properties in question, they must share some form of observational evidentiary basis. Here we propose what we take to be a minimal, pragmatic approach to the justification of the existence of properties, based on our ability to identify their modes of interaction with their environments.[52]

A core aspect of the justification for a claim about the existence of a property is, simply, the observation that an object interacts with its environment in particular ways. The term "interaction" can itself be interpreted in a variety of ways, but in the context of measurement science (see in particular Sects. 2.3, 3.1, and 4.3.4), a starting point is the observation of what we have referred to as a transduction effect, i.e., an empirical process that produces variations of a (response, effect, output) property as effects of variations of one or more (stimulus, cause, input) properties.

One could argue that an even earlier starting point is simply the observation of variation in some empirical phenomenon (event, process, etc.). If we may help ourselves to the assumption that there are no uncaused events (at least not on a sufficiently broad conception of causality; for general discussions, see, e.g., Beebee, Hitchcock, & Menzies, 2009; see also Markus & Borsboom, 2013), then from the observation of an event one may infer the existence of causal influences, though of course one may initially know little or nothing about the nature of these influences.[53] Progressively, through empirical interaction with relevant phenomena, we may arrive at a state of knowledge and technology such that a transduction effect can be dependably reproduced under specified conditions, which brings us back to the "starting point" referenced in the previous paragraph. Such a transduction effect may become the basis of a direct method of measurement (see Sect. 7.3): through the calibration of the transducer, the values of the output property (i.e., the instrument indication) are functionally related to values of the input property (i.e., the property under measurement). For example, temperatures can be measured by means of differences of expansion of mercury in a glass tube, and reading comprehension abilities can be measured by means of differences in patterns of responses to questions about a particular text. Such cases presuppose the observability of a property Y (e.g., shape, color, pattern of responses to test questions), whose differences are accounted for as being causally dependent on differences in the property under consideration P, via an inference of the kind $Y = f(P)$, where f is the function that models the cause-effect relation: the property P is the cause of observed changes of Y, and therefore it exists.

All this said, if a property P is *only* known as the cause of observable effects in the context of a single empirical situation (experimental setup etc.) – that is, if there is only a single known transduction effect of which instances of P are the input, and where the transduction itself is understood only at a black-box level – then knowledge of P is obviously highly limited; such a situation might be associated with an operationalist perspective on measurement, and would thus

[52] To be clear, we do not aim to provide a *sufficient* set of criteria for the justification of a claim about the existence of any given property, as this would surely involve issues specific to that property.

[53] Our stance here is broadly consistent with Ian Hacking's (1983) perspective on entity realism, which entails that a claim about the existence of an entity is justified if it can be used to create effects that can be investigated and understood independently of their cause. As Hacking famously put it, in reference to experiments involving the spraying of electrons and positrons onto a superconducting metal sphere: "if you can spray them, then they are real" (p. 24). To this we would add a friendly amendment: if you can spray them, *something* is real, though it remains an open question to what extent the actual causal forces at work are well-described by our current best theories and terminology. This is easily illustrated by the historical example of phlogiston (as also discussed in Box 5.1): although contemporary theories deny the existence of the substance referred to as "phlogiston" by 17th- and 18th-century theorists, contemporary theories would not deny the existence of the causal forces responsible for the putative effects of phlogiston (e.g., flammability, oxidation, rusting), but instead offer more nuanced explanations for the identity and mechanisms of these causal forces.

inherit the limitations of that perspective (see Sect. 4.2.2), or might simply be a very early phase in the identification of an empirical property, setting the stage for investigations of the causal relevance of the property in situations other than this single transduction effect. Indeed, in general, absent the availability of multiple, independent sources of knowledge about P, in particular about its role in networks of relationships with other phenomena (properties, outcomes, events, etc.), knowledge about P might be considered vacuous or trivial.

For example, a claim about the existence of hardness as a property of physical objects can be justified in a simple way by the observation that one object scratches another: hardness (P) is what causes (f) observable scratches (Y) to appear given an appropriate experimental setup. Were this the *only* source of knowledge about hardness, the correct name for P would arguably be something like "the property that causes the effect Y", e.g., "the ability to produce scratches", rather than a label as semantically rich as "hardness". But, of course, this is not the only way in which hardness is known: even simple lived experience can corroborate our common sense about the ways in which objects made of different materials interact with one another; this is further corroborated by alternative methods for measuring hardness such as via observation of indentations under specified conditions. In other words, we have access to knowledge about the property of hardness also independently of that particular cause-effect relationship f, and this knowledge is consistent with what f models as the cause of Y. This shows that the procedure of checking which objects scratch which other objects does not *define* hardness, but instead may become a method for *evaluating* it.[54]

Thus, as investigations reveal functional relations connecting P to multiple phenomena (properties, outcomes, events, etc.) whose existence can be assessed independently of such relations, P becomes part of a system of interrelated properties, sometimes called a *nomic network*.[55] The identification of such relations (referred to in the VIM as a "set of quantities [or more generally, properties] together with a set of noncontradictory equations relating those quantities", JCGM, 2012: 1.3) is important not only because it expands the explanatory and predictive value of knowledge of P,[56] but also for two additional reasons specifically related to measurement. The first is that such knowledge may suggest alternative methods for directly measuring a given property: for example, temperatures could also be measured by means of differences of electric potential via the thermoelectric effect, and reading comprehension abilities could also be measured by observing how well an individual is able to carry out a set of instructions after having read a relevant text. This corresponds to the minimal example of a nomic network as shown in Fig. 6.9, in which the three properties P, Y, and Z are connected via the two functions $Y = f(P)$ and $Z = g(P)$.[57] The causal relationship between P and either Y or Z – or both – could be used as the basis for a direct measurement of P. This kind of relationship of Y and Z to P is referred to as *reflective* in the context of latent variable modeling (see, e.g., Edwards & Bagozzi, 2000).

[54] The same reasoning applies to the case of educational tests, which would in general not be valued unless the competencies they purport to measure are demonstrably valuable in contexts beyond the immediate testing situation. For further arguments along these lines, see also Rozeboom (1984).

[55] The adjective "nomic" comes from the ancient Greek νόμος, "nomos", meaning <law>. When attributed to a conceptual network it refers to a set of entities (in this case general properties) interconnected via relations interpreted as laws. The paradigmatic example of this is the International System of Quantities (ISQ), a system of (general) quantities based on length, mass, duration, intensity of electric current, thermodynamic temperature, amount of substance, and luminous intensity (JCGM, 2012: 1.6), from which other physical quantities may be derived through physical laws.

[56] In this we agree with Carl Hempel: "We want to permit, and indeed count on, the possibility that [candidate properties] may enter into further general principles, which will connect them with additional variables and will thus provide new criteria of application for them" (1952: p. 29).

[57] Y and Z would be expected to covary as the effects of the common cause P. This is, in fact, the canonical example of how "correlation is not causation": the observation that two properties Y and Z correlate may be explained by the presence of a third, "hidden" property P which is their common cause.

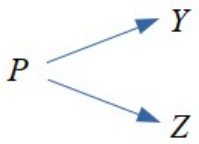


Fig. 6.9 A simple nomic network laying the groundwork for the direct measurement of *P* through multiple means (where $P \rightarrow Y$ means that *P* is the cause of *Y*)

The second measurement-related reason for the importance of knowledge of such functional relations is that they may become the basis for *indirect* methods of measurement (see Sect. 7.2), in which the results of prior direct measurements are used as input properties for the computation of a value of the output property (i.e., the measurand), as, for example, when densities are measured by computing ratios of measured values of masses and volumes. Here the property *P* whose existence is questioned is a function of other properties, say *Y* and *Z*, whose existence is already accepted, as depicted in Fig. 6.10. This kind of relationship of *Y* and *Z* to *P* is referred to as *formative* in the context of latent variable modeling (again see Edwards & Bagozzi, 2000).

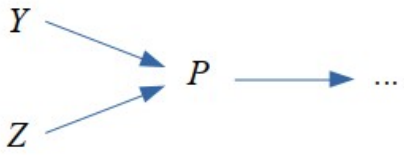


Fig. 6.10 A simple example of a nomic network laying the groundwork for the indirect measurement of *P*

A clarification is in order on this matter: if a property is *only* known through a single function of other properties, in which case the functional relation $P = f(Y, Z)$ would serve as the *definition* of a previously unknown entity *P*, there would be no basis for claiming that *P* is an independently-existing empirical property; rather, what is calculated by *f* would simply be a variable that summarizes (some of) the available information about the properties *Y* and *Z* (as, again, is the case for hage, defined as the product of the height and age of a human being; Ellis, 1968: p. 31). Summaries can, of course, have substantial utility, but as per the previous discussion of the distinction between empirical properties and mathematical variables, mathematical creativity is in itself insufficient for the generation of new empirical properties. As before, it is the availability of independent sources of knowledge about the property in question that lends credence and importance to claims regarding its existence, as is the case with force: although $F = ma$ may be considered to be a definition of force, there are in fact means of knowing force independently of (but consistent with) Newton's second principle, as, for example, Coulomb's law, which connects force to quantity of electric charge.

In sum, our approach to the justification of claims about the existence of properties is consistent with the philosophical perspective sketched in Sect. 4.5, which we described as *pragmatic realism* or *model-based realism*. The approach is realist, insofar as it focuses on justification for claims regarding the existence of empirical properties, and by so doing helps clarify the distinction between empirical properties and mathematical variables, and more generally the interface between the empirical world and the informational world; this also helps set the stage for a clear distinction between measurement and computation, discussed further in the following chapter. The approach is pragmatic, insofar as the emphasis of the proposed criteria for evaluating our beliefs about the existence of properties is on the practical *consequences* of those beliefs; this is consistent with the familiar refrain of pragmatic philosophers that "a difference that makes no difference is no difference", or, as put more specifically

by Heil, “a property that made no difference to the causal powers of its possessors would, it seems, be a property the presence of which made no difference at all” (2003: p. 77). Finally, the approach is model-based, insofar as the role of models (of general properties, measurands, environments, and the measurement process) is given primacy: this is the topic to which the following chapter is devoted.

References

Babbie E. (2013). *The practice of social research* (13th ed.). Belmont: Wadsworth.

Beebee, H., Hitchcock, C., & Menzies, P. (Eds.) (2009). *The Oxford handbook of causation*. Oxford: Oxford University Press.

Bentley, J. P. (2005). *Principles of measurement systems* (4th ed.). New York: Pearson.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71(3)*, 425–440.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch model. *Measurement, 146*, 961–971.

Bunge, M. (1973). On confusing ‘measure’ with ‘measurement’ in the methodology of behavioral science. In M. Bunge (Ed.), *The Methodological unity of science* (pp. 105–122). Dordrecht: Reidel.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press.

Carnap, R. (1966). *Philosophical foundations of physics*. New York: Basic Books.

Chang, H. (2001). How to take realism beyond foot-stamping. *Philosophy, 76(295)*, 5–30.

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

Chrisman, N. R. (1998). Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems, 25*, 231–242.

De Boer, J. (1995). On the history of quantity calculus and the International System. *Metrologia, 31*, 405–429.

Doebelin, E. (1966). *Measurement systems: Application and design* (5th ed. 2003). New York: McGraw-Hill.

Dybkaer, R. (2013). Concept system on ‘quantity’: formation and terminology. *Accreditation and Quality Assurance, 18*, 253–260.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155–174.

Eisenhart, C. (1963). Realistic evaluation of the precision and accuracy of instrument calibration systems. *Journal of Research of the National Bureau of Standards. Engineering and Instrumentation*, *67C(2)*. Retrieved from nvlpubs.nist.gov/nistpubs/jres/67C/jresv67Cn2p161_A1b.pdf

Ellis, B. (1968). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

Flynn, J. R. (2009). *What is intelligence: Beyond the Flynn Effect*. Cambridge: Cambridge University Press.

Freund, R. (2019). *Rasch and Rationality: Scale typologies as applied to Item Response Theory*. Unpublished doctoral dissertation, University of California, Berkeley.

Giordani, A., & Mari, L. (2012). Property evaluation types. *Measurement, 45*, 437–452.

Girard, G. (1994). The third periodic verification of national prototypes of the kilogram (1988-1992). *Metrologia, 31(4)*, 317–336.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 510–522.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Heil, J. (2003). *From an ontological point of view*. Oxford: Clarendon Press.

Hempel, C. G. (1952). *Fundamentals of concept formation in physical science*. Chicago: Chicago University Press.

Holder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse, 53*, 1–46. Transl. J. Michell, C. Ernst, The axioms of quantity and the theory of measurement, Journal of Mathematical Psychology, 40, 235–252, 1996.

Holland, P. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577–601.

International Bureau of Weights and Measures (2019). *The International System of Units (SI)* (9th ed.). Sèvres: BIPM. Retrieved from www.bipm.org/en/si/si_brochure

International Organization for Standardization (ISO) and other three International Organizations (1984). *International vocabulary of basic and general terms in metrology (VIM)* (1st ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML).

Joint Committee for Guides in Metrology (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.). Sèvres: JCGM (2008 version with minor corrections). Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Kelly, E. J. (1916). The Kansas silent reading tests. *Journal of Educational Psychology, 7*, 63–80.

Kim, J. (1998). *Mind in a physical world*. Cambridge: MIT Press.

Kintsch, W. (2004) The Construction-Integration model of text comprehension and its implications for instruction. In R. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed.). Newark, DE: International Reading Association.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (vol. 1). New York: Academic Press.

Kuhn, T. S. (1969). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kyburg Jr, H. E. (1984). *Theory and measurement*. Cambridge: Cambridge University Press.

Lodge, A. (1888, July). The multiplication and division of concrete quantities. *Nature, 38*, 281–283. Retrieved from www.nature.com/articles/038281a0

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750–751.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research, 31(4)*, 603–616.

Mari, L., & Giordani, A. (2012). Quantity and quantity value. *Metrologia, 49*, 756–764.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Mari, L., Maul, A., & Wilson, M. (2018). On the existence of general properties as a problem of measurement science. *Journal of Physics: Conference Series 1065*, 072021.

Mari, L., & Sartori, S. (2007). A Relational Theory of Measurement: traceability as a solution to the non-transitivity of measurement results. *Measurement, 40*, 233–242.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15*, 51–69.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement, 131*, 764–770.

Maurin, A. S. (2018). Tropes. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/tropesMcGrane, J., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement, 152*. doi.org/10.1016/j.measurement.2019.107346

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale: Erlbaum.

Michell, J. (1999). *Measurement in psychology – Critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10(5)*, 639–667.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology, 14(1)*, 121.

Michell, J. (2005). The logic of measurement: a realist overview. *Measurement, 38*, 285–294.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge.

Mundy, B. (1987). The metaphysics of quantity. *Philosophical Studies, 51*, 29–54.

Narens. L. (1985). *Abstract measurement theory*. Cambridge: MIT Press.

Narens. L. (2002). *Theories of meaningfulness*. London: Lawrence Erlbaum Associates.

Popper, K. (1959). *The logic of scientific discovery*. Abingdon-on-Thames: Routledge.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Roberts, F. S. (1979). *Measurement theory with applications to decision-making, utility and the social sciences*. Reading, MA: Addison-Wesley.

Rossi, G. B. (2006). A probabilistic theory of measurement. *Measurement, 39*, 34–50.

Rossi, G. B. (2007). Measurability. *Measurement, 40*, 545–562.

Rozeboom, W. W. (1984). Dispositions do explain: Picking up the pieces after hurricane Walter. In J. R. Royce & L. P. Mos (Eds.), *Annals of theoretical psychology* (Vol. 1, pp. 205–224). . New York: Plenum.

Russell, B. (1903). *The principles of mathematics*. London: Bradford & Dickens.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge: MIT Press.

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford: Oxford University Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423 and 623–656.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A, 42*, 509–524.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 667–680.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In C. West Churchman & P. Ratoosh (Eds.), *Measurement, definitions and theories* (pp. 18–63). Wiley, New York.

Suppes, P. (1951). A set of independent axioms for extensive quantities. *Portugaliae Mathematica, 10(4)*, 163–172.

Suppes, P., & Zanotti, M. (1992). Qualitative axioms for random-variable representation of extensive quantities. In C. W. Savage & P. Ehrlich (Eds.), *Philosophical and foundational issues in measurement theory* (pp. 39–52). Hillsdale, NJ: Lawrence Erlbaum.

Tal, E. (2019). Individuating quantities. *Philosophical Studies, 176(4)*, 853–878.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–544.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*, 65–72.

von Helmholtz, H. (1887). *Zählen und Messen Erkenntnis–theoretisch betrachtet, Philosophische Aufsätze Eduard Zeller gewidmet*. Leipzig: Fuess. Translated as: Numbering and measuring from an epistemological viewpoint. In W. Ewald (Ed.), From Kant to Hilbert: A sourcebook in the foundations of mathematics: Vol. 2 (pp. 727–752). Oxford: Clarendon Press, 1996.

Wetzel, L. (2018). Types and tokens. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/types-tokens

Weyl, H. (1949). *Philosophy of mathematics and natural science*. Princeton: Princeton University Press.

Wilson, M., Mari L,. & Maul, A. (2019). The status of the concept of reference object in measurement in the human sciences compared to the physical sciences. Proc. Joint International IMEKO TC1+ TC7+ TC13+TC18 Symposium, St Petersburg, Russian Federation, 2–5 July 2019, IOP Journal of Physics: Conference Series, 1379, 012025. Retrieved from iopscience.iop.org/article/10.1088/1742-6596/1379/1/012025/pdf

Chapter 7.
Modeling measurement and its quality

This chapter aims to propose a general model of a measurement process, consistent with the ontological and epistemological commitments developed in the previous chapters. Starting from the premise that any measurement is grounded on an empirical process, we propose a characterization of measurement methods related to the complementary roles of empirical and informational components, by broadly distinguishing between direct and indirect methods of measurement, where indirect measurements necessarily include at least one direct measurement. The stages of direct measurement are then analyzed and exemplified in reference to both physical and psychosocial properties. This is the basis for discussing once again the quality of measurement, now described in terms of the complementary requirements of object-relatedness ("objectivity") and subject-independence ("intersubjectivity").

7.1 Introduction

Despite – or perhaps, to at least some extent, because of – the ubiquity of measurement-related concepts and discourse, there remains a remarkable lack of shared understanding of these concepts across (and often within) different fields, perhaps most visibly reflected in the vast array of distinct proposed definitions of <measurement> itself, as discussed in Sect. 4.2. It would seem, then, that the clarification of foundational measurement concepts should (continue to) to be a high priority, not only in terms of the definition of <measurement>, but also of the identification of those features of measurement that justify its epistemic authority, i.e., its commonly-afforded degree of public trust and social prestige: the claim that "measurement is often considered a hallmark of the scientific enterprise and a privileged source of knowledge relative to qualitative modes of inquiry", in Eran Tal's words (2020). Justification of the epistemic authority of measurement and its results, in turn, depends on identifying those features of the measurement process that ensure (or, at least, confer high likelihood upon) the quality of its results. We argue that *these features are independent of the domain of application*, and thus in principle apply equally to the measurement of physical and psychosocial properties; as such, this topic is a key component of our endeavor toward a conceptual framework of measurement across the sciences.

As described in Chap. 4, since the second half of the 20th century, scholarly treatment of the foundational aspects of measurement has largely focused on mathematical criteria rather than the concrete realization of the process, as exemplified by claims such as that "we are not interested in a measuring apparatus and in the interaction between the apparatus and the objects being measured. Rather, we attempt to describe how to put measurement on a firm, well-defined foundation" (Roberts, 1979: p. 3) and "the theory of measurement is difficult enough without bringing in the theory of making measurements" (Kyburg, 1984: p. 7). This emphasis on formal characterizations of measurement may be in part explained by the expansion of measurement into new domains and the related need to abandon characterizations and requirements that were needlessly tied to specific areas. In particular, it is clear that the measurement of non-physical properties cannot conform to expectations based on the traditional use of measuring instruments operating on the basis of transduction effects implemented by physical sensors. As a consequence, interpretations of

measurement have become so abstract that they may be unable to provide a convincing and useful demarcation of measurement from formally similar processes that are generally thought to lack epistemic authority, such as most statements of subjective judgments and opinions.

Our position on this matter is pragmatic: there is a social interest in sharing a scientific and technical concept system across disciplines,[1] particularly in the case of an infrastructural activity like measurement, and there is a social acknowledgment of the epistemic authority of measurement, which has critical consequences in particular in terms of public trust attributed to the outcomes of putative measurement processes and the resources devoted to such processes. If the claim that a given evaluation process is a measurement could be invoked at will, without understanding or concern for what has historically made it a valued practice, measurement itself would become simply a rhetorical device, risking the discredit of its practice in general.

In Sect. 4.4.2 we presented our claim that measurement is most appropriately characterized by empirical rather than mathematical conditions, as grounded on a model-dependent realism, introduced in Sect. 4.5 and developed in Chaps. 5 and 6, about the objects of measurement, i.e., properties. We develop here that claim by proposing that *measurement is a process characterized by its structure*, not only by the specification of the relationship connecting its inputs to its outputs: what is required for justifying the dependability of measurement results is *the specification of how the process does what it does*. Whereas an input-output relationship relies solely on a black box model, a structural model involves identification of the invariant aspects of the empirical process, and therefore looks "inside the black box". And this, we argue, is what provides justification for the claim that measurement results are publicly trustworthy. As a corollary, any purely black box model cannot adequately account for the all relevant features of measurement, and thus is not sufficient for the purpose of understanding the quality of measurement results.

As already highlighted, the conditions presented in Chap. 2 – i.e., that measurement (i) is both an empirical and an informational process (ii) designed on purpose, (iii) whose input is an empirical property of an object, and that (iv) produces information in the form of values of that property – are necessary but not sufficient for a process to be considered a measurement. We propose here that the missing sufficient conditions are provided by *a structural model of the process of measuring*. As a corollary, such an integrated set of necessary and sufficient conditions provides a characterization of measurability: *a property is measurable if and only if there exists a property-related process fulfilling these conditions*.

From this perspective, the analysis of the structure of a measurement process plays a crucial role: in the metrological tradition, the general description of the structure of such a process is provided by a so-called *measurement method*, defined as a "generic description of a logical organization of operations used in a measurement" by the *International Vocabulary of Metrology* (VIM) (JCGM, 2012: 2.5). A key related distinction is between direct and indirect (methods of) measurement, first introduced in Sect. 2.3. To this we first devote our attention here, using the model proposed by Giordani and Mari (2019) as a starting point, which we develop to encompass the scenarios that arise when the model is applied across the sciences.

[1] A basic reason for the complexity of this endeavor is the (usually unavoidable and in fact appropriate) specialization of the scientific and technical disciplines, which triggers the construction of specific terminologies. An interesting example of an attempt to overcome lexical hyper-specialization while maintaining scientific and technical correctness is Electropedia, "the world's most comprehensive online terminology database on "electrotechnology", containing more than 22000 terminological entries [...] organized by subject area" (Electropedia makes the series of standards IEC 60050 freely accessible online at www.electropedia.org).

7.2 Direct and indirect measurement

Though it contradicts what is currently specified by the VIM, which defines <measurement> to be an experimental process (JCGM, 2012: 2.1), and also against our own presentation of this as a necessary condition (see Sect. 2.2.1), the idea that measurement is not necessarily empirical is not new. In his seminal 1920 book, Norman Campbell defined <measurement> as "the process of assigning numbers to represent qualities" (1920: p. 267). A linguistic detail is again revealing: Campbell wrote "*the* process", not "*a* process", thus supposedly implying that *any* process of assigning numbers to properties – with only a slight paraphrase – is a measurement. This was conceived in the context of a foundationalist endeavor aimed at framing measurement as a core enabler of science (1920: p. 267):

> Physics could be distinguished from other sciences by the part played in it by measurement. Other sciences measure some of the properties they investigate but it is generally recognized that when they make such measurements they are always depending, directly or indirectly, on the results of physics. All fundamental measurements belong to physics, which might almost be described as the science of measurement.

Here the term "fundamental measurement" is used with a specific meaning, introduced by Campbell himself: being fundamental is what characterizes properties whose instances can be directly compared with each other by equivalence and order and can be additively combined (Campbell used the term "physical addition", p. 279, for what has later been called "concatenation", e.g., by Krantz et al., 1971: p. 2). This fundamentality is due to the fact that, for any three objects a, b, and c having the property P the construction introduced in Sect. 6.3 applies, i.e., if $P[a] \approx P[b]$ and $P[a] + P[b] \approx P[c]$ then $P[c] = 2\ P[a]$, or $P[c] / P[a] = 2$: with no other conditions – hence with no previously defined units, measurement standards, metrological traceability chains, instrument calibrations, etc. – numbers have been thus assigned to ratios of properties. Moreover, were $P[a]$ conventionally set as the unit u, the previous equality would become $P[c] = 2$ u, which is an example of a Basic Evaluation Equation and the simplest case of a measurement result, under the condition that measurement uncertainty need not to be reported explicitly. If this is all that is required for (fundamental) measurement, then only the adjective "physical" in the term "physical addition" ties measurement to the empirical world: indeed, exactly the same structural conditions may apply to properties of mathematical objects, through purely computational processes (and this could be indeed the underlying reason for the NIST definition of <measurement> as a "an experimental *or computational* process", as quoted in Footnote 11 of Chap. 5).

In fact, given his interest in establishing measurement as a foundation for science, Campbell included an almost incidental second condition: "in order that a property should be measured as a fundamental magnitude, involving the measurement of no other property, it is necessary that a physical process of addition should be found for it" (p. 267). Together with additivity, he considered "involving the measurement of no other property" as the basis for the distinction between fundamental and derived properties, and then between fundamental and derived measurement. This distinction was later refined and became a sort of default reference, in particular in the version presented by Brian Ellis (1968), where measurements – and more properly measurement methods – are classified as either fundamental (for which Ellis also used the term "direct") or indirect, so that every measurement method that is not fundamental / direct is considered to be indirect.

As the scope of measurement broadened, and (especially) psychophysicists and psychologists argued against the necessity of physical addition operations for measurement, Campbell's second

condition – that no other properties are involved in the process – became the characterizing feature of those methods which were then called "direct methods" of measurement, as exemplified by the first edition of the VIM, which defines <direct method of measurement> as a "method of measurement in which the value of a measurand is obtained directly, rather than by measurement of other quantities functionally related to the measurand" and <indirect method of measurement> as "method of measurement in which the value of a measurand is obtained by measurement of other quantities functionally related to the measurand" (ISO et al., 1984: 2.13 and 2.14). Similar definitions or characterizations can be found elsewhere in the literature, for example (Lira, 2002: p. 39)

> Many times the information about the measurand is acquired from the readings of a single measuring instrument. We will refer to such a quantity as subject to a *direct* measurement. However, the metrological meaning of the word 'measurement' is more ample: it also refers to quantities whose values are *indirectly* estimated on the basis of the values of other quantities, which in turn may or may not have been directly measured.

or (Gertsbakh, 2003: p. viii)

> There are two principal types of measurements: direct and indirect. For example, measuring the voltage by a digital voltmeter can be viewed as a direct measurement: the scale reading of the instrument gives the desired result. On the other hand, when we want to measure the specific weight of some material, there is no such device whose reading would give the desired result. Instead, we have to measure the weight W and the volume V, and express the specific weight p as their ratio: $p = W/V$. This is an example of an indirect measurement.

The idea that a measurement is performed according to a direct method, is direct for short,[2] if it does not involve properties other than the measurand seems to be accepted also by the *Guide to the expression of uncertainty in measurement* (GUM), which notes that "in most cases, a measurand *Y is not measured directly*, but is determined from N other quantities X_1, X_2, ..., X_N through a functional relationship f, $Y = f(X_1, X_2, ..., X_N)$" (JCGM, 2008: 4.1.1; emphasis added). This was discussed by Walter Bich as follows (2008: p. 272; emphasis added):

> even the simplest, *seemingly direct measurements* [...] fall into this categorization. For example, the indication of a bathroom balance, which is expressed in divisions of the scale, is not the measurand Y (which is the mass of the person in kilograms), but simply one of the input quantities, say, X_1. The measurand is obtained from the indication X_1, perhaps repeated two or three times, and a series of corrections $X_2, X_3, ..., X_N$ (the zero and the span of the scale, and perhaps its linearity, or the deviation of the local acceleration due to gravity from that of the place in which the balance was manufactured and adjusted).

The consequence is then straightforward: from this perspective, since "even the simplest model will be incomplete if corrections to the indications of the instruments used in direct measurements are not taken into account, [...] *no measurement can strictly be considered to be 'direct'*." (Lira, 2002: p. 50; emphasis added).

Something peculiar can be observed in this sequence of apparently coherent steps: *it started by emphasizing that the foundational role of measurement is guaranteed by fundamental, or direct,*

[2] About direct and indirect methods of measurement, see also the discussion in Boumans (2007: 9.3).

measurement, and ended with the admission that, in practice, no measurement can be, in this sense, direct!

This shift was plausibly driven by the recognition that even in the simplest measurements some computational activity is required, as part of the modeling of the empirical process that takes place in the interaction between the object under measurement and the measuring instrument in the given experimental context. And in fact, significantly, the functional relationship f mentioned above is described by the GUM as a *model of measurement* (JCGM, 2008: 4.1.2), thus with the understanding that "observations are never interpreted independently of some abstract model of the [...] system" (Cook, 1994: p. 4), because "observation of x is shaped by prior knowledge of x" (Hanson, 1958: p. 19), i.e., all observations are unavoidably "theory-laden".[3]

Let us take for granted that measurement cannot produce "pure empirical data" – whatever this could mean – both because some background model is always and unavoidably, though sometimes only implicitly, present also in measurement, and because measurement produces information, not empirical entities, and therefore it must include an informational stage (see Sect. 2.3 for a preliminary justification of this). However, this only affects the claim that there can be methods of measurement that exclusively require empirical operations, a position that in fact can be then easily ascertained as wrong. What is at stake here is way more than that: *it is the very possibility of providing a justification for the epistemic authority of measurement, as based on an understanding of the relationship that measurement establishes between the empirical and the information world*.

Indeed, the acknowledged foundational role of measurement for empirical sciences needs to be explained by showing where, in principle, measurement ends and other information production processes (such as computation, opinion making, and guess) start. The thesis proposed and developed here is that this justification must be found in the distinction between direct and indirect (methods of) measurement, thus providing a way to overcome this possible crisis of foundationalism in measurement (Mari, 2005).

Consider again the sentence quoted above from the GUM: the functional relationship f, $Y = f(X_1, X_2, ..., X_N)$, through which a value is attributed to the measurand Y is what the VIM calls a *measurement model*, i.e., a "mathematical relation among all quantities known to be involved in a measurement" (JCGM, 2012: 2.48). This assumes the quantities X_i to be somehow evaluated, but does not set any constraint on this matter: in particular, could their values come from simulations, or hypotheses, or guesses? Without an independent characterization of the concept <measurement>, the idea that a mathematical entity f is a *model* of a measurement, and therefore an interpretation of what is a measurement, blurs everything. And the issue is not settled simply by acknowledging the unavoidable presence of models in measurement. Indeed, as we discuss below, together with what the VIM calls "measurement models" there are at least two other kinds of models that are relevant in this context:

- *models of the measuring instrument behavior*, with respect to the environmental properties that influence the relation between the property being measured and the instrument indication;[4]

[3] This is possibly the reason for the statement in the GUM-related document about conformity assessment that "in a typical measurement, the measurand of interest is not itself observable. The length of a steel gauge block, for example, cannot be directly observed, but one could observe the indication of a micrometer with its anvils in contact with the ends of the block." (JCGM, 2012: p. vii). However, one could point out that the observation of the indication of a measuring instrument is an empirical observation in turn: shouldn't this impossibility recursively apply then?

[4] This is related to what is referred to as the measurement model part of a structural equation model (SEM), or a *measurement model* for short in the context of Structural Equation Modeling (SEM; see, e.g., Skrondal & Rabe-Hesketh, 2004). SEMs are a popular class of statistical models used in the human sciences, although such models have a number of different purposes.

- *models of the measurand*, with respect to the way the measurand is affected by, and more generally related to, other properties.[5]

An exploration of the distinction between these (kinds of) models and their relations will lead us to a better understanding of the role of models both *of* and *in* measurement, and in consequence of measurement as such. The position that we are going to propose here is that what the VIM calls "measurement model" is not a model of a measurement, but a component of an evaluation process, that is an indirect measurement only if it includes also at least one direct measurement: hence, the characterization of direct measurements is pivotal for our endeavor.

7.2.1 Recovering a meaningful distinction between direct and indirect measurement

The term "direct measurement" does not have a single inherent meaning and it is not trademarked, of course: to our knowledge Campbell did not use it; Albert de Forest Palmer characterized it as "the determination of [the value] by direct observation of the measured quantity, with the aim of a divided scale or other indicating device graduated in terms of the chosen unit" (1912: p. 11), and Ellis might have been the first to include it in a structured analysis of measurement, though with the not-so-clear characterization that "direct measurement is any form of measurement which does not depend upon prior measurement" (1968: p. 56). Were the idea of direct measurement intended as designating a model-free form of measurement, it should simply be dismissed, given the acknowledgment that purely empirical and model-free measurement is not possible.

Rather, we propose to recover a meaningful distinction between direct and indirect measurement by restarting from the definition in IEC 60050 / Electropedia of <direct (method of) measurement>: "method of measurement in which the value of a measurand is obtained directly, without the necessity for supplementary calculations based on a functional relationship between the measurand and other quantities actually measured" (IEC: 311-02-01). Admittedly, the definition is circular, since <direct measurement> is defined in terms of directly obtaining something. A possible amendment of the IEC definition that eliminates this circularity could be as follows:

> (preliminary, provisional characterization of <direct method of measurement>) *a measurement method is direct if it is based on the use of a measuring instrument that is designed to empirically interact with properties of the same kind as the measurand*

where the adjective "direct" refers to such a condition of direct interaction.[6] It is also worth noting that Ellis acknowledged the importance of this typology and called it "associative measurement", though, in contrast to the proposal we advance here, he considered it a form of non-fundamental and therefore indirect measurement (1968: p. 90).

[5] This is related to what is referred to as the structural model part of the structural equation model (SEM), or *structural model* for short.

[6] The distinction between direct measurements and processes of evaluation which are not direct is sometimes accepted as unproblematic. For example, consider this quotation from John Taylor (1997: p. 45): "Most physical quantities usually cannot be measured in a single direct measurement but are instead found in two distinct steps. First, we *measure* one or more quantities that can be measured directly and from which the quantity of interest can be *calculated*. Second, we use the measured values of these quantities to calculate the quantity of interest itself. For example, to *find* the area of a rectangle, you actually *measure* its length l and height h and then *calculate* its area A as $A = l\,h$. Similarly, the most obvious way to *find* the velocity v of an object is to *measure* the distance traveled, d, and the time taken, t, and then to *calculate* v as $v = d/t$." (all emphases added). Interestingly, the process of indirect evaluation is not even called a measurement here, as if only direct measurements were measurements, in opposition to Lira's conclusion mentioned above, that "no measurement can strictly be considered to be 'direct'".

Many measurements are direct in this sense: basically, whenever the core empirical process is a transduction from a property of the same kind as the measurand to another property (which in current physical applications is typically an electric quantity), as implemented in a sensor. Examples of direct measurement would then include both the measurement of temperature by means of an alcohol thermometer – where the thermometer interacts with the object under measurement and transduces its temperature into a position of the upper surface of the alcohol in the glass tube (temperature has well-known characteristics today, yet has a long and complex measurement history; see for example Chang, 2007) – and the measurement of reading comprehension ability by means of a test – where the items in the test interact with the reader and transduce her reading ability into a pattern of item responses.

However, we consider the characterization above provisional because it provides at most a necessary but not sufficient condition for a method to be direct. In fact, it turns out that a measurement could be indirect – according to the usual way the adjective is used as attributed to measurement methods – even if the property with which the measuring instrument interacts and the measurand are of the same kind.

The philosophical tradition gives us a nice example of this. When visiting some Egyptian priests, Thales asked them about the height of the Great Pyramid and, since they were unable to answer the question, he devised a procedure to get the information by himself. Thales measured the length of the shadow cast by the Pyramid and compared it with the length of the shadow cast by a vertical rod of known length, so that he could use his theorem on the proportionality of the sides of similar triangles to infer the Pyramid's height. If asked to assess this case, we would say that Thales performed an *indirect* measurement, through the direct measurement of the length of the shadow cast by the Pyramid and the length of the shadow cast by the rod, though these quantities are of the same kind as the measurand, being in fact all lengths. Therefore, the previous characterization may be refined as follows:

> (refined, provisional characterization of <direct method of measurement>) *a measurement method is direct if it is based on the use of a measuring instrument that is designed to empirically interact with properties of the same kind as the measurand and is actually coupled with the object carrying the measurand*

In Thales' case, he needed to measure first the lengths of two shadows: while the instrument he used to this purpose was designed to empirically interact with lengths, it was not coupled with the object carrying the measurand, i.e., the Pyramid, but with another objects, i.e., the shadows of the Pyramid and of the rod.

While again still provisional, this characterization allows us to describe the basic structure of a direct measurement process, as it is presented in Sect. 2.3, as follows.

- *Transduction*. The measuring instrument is put in interaction with the object under measurement with respect to a property of the object; as a result, the instrument changes its state, by transducing the property under measurement to another property, i.e., the instrument indication.
- *Instrument scale application*. The instrument indication, which is still an empirical property, is associated with an indication value through the application of the instrument-related scale; this is the crucial step in which an empirical entity (e.g., the position of the upper surface of the alcohol in the tube of a thermometer; a pattern of responses to a set of test items) is associated with an information entity (e.g., a value of position; a number of correct answers).
- *Calibration function computation*. The indication value is mapped to a measurand value by computing the instrument calibration function, consistently with the VIM definition, according

to which the second step of a calibration "establish[es] a relation for obtaining a measurement result from an indication" (JCGM, 2012: 2.39), and where corrections to the indications of the instruments are part of the calibration function.

Hence, the sequence

transduction → instrument scale application → calibration function computation

may be interpreted as a process mapping a property of an object to a measured value, as depicted in Fig. 7.1 (measurement uncertainty is not considered here).

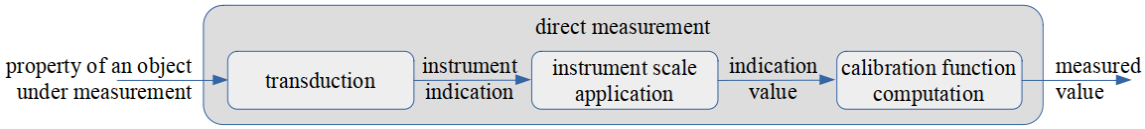


Fig. 7.1 The basic structure of a direct measurement (adapted from Fig. 2.10)

It is crucial that both empirical and non-empirical stages – i.e., transduction and calibration function computation respectively – are required to complete this process. This provides a key to distinguishing measurement from computation and to highlighting a crucial asymmetry with respect to their sources of justification: a necessary, though insufficient, condition for computation to produce justified results is that its input data are justified. Hence, if such input data comes from and is about empirical properties, only procedures based on direct measurement provide empirically justified results. Even in the case of a measurement which is non-direct, one or more measuring instruments that produce the input for the computations are required to empirically interact with properties of the relevant objects. This allows us to propose as a provisional characterization that

> (provisional characterization of <indirect method of measurement>) *a measurement method is indirect if it is not direct, i.e., if the measuring instruments are not designed to empirically interact with properties of the same kind as the measurand or if they are not coupled with the object carrying the measurand*

As a consequence, any measurement requires that at least one direct method be applied, for the measurement of one or more "intermediate" measurands, as in the case of the (indirect) measurement of density via the (direct) measurements of mass and volume, which then operate as intermediate measurands. This is consistent with the definition in IEC 60050 / Electropedia of <indirect (method of) measurement> as a "method of measurement in which the value of a quantity is obtained from measurements made by direct methods of measurement of other quantities linked to the measurand by a known relationship" (IEC: 311-02-02). If such direct measurements are considered as black boxes, which produce values of the n–1 intermediate measurands, the functional relationship f presented by the GUM is the combination function of an indirect measurement, which models the relation among some properties of the object under measurement (density, mass, and volume in the example), but not the behavior of an instrument. Fig. 7.2 summarizes this conceptual structure.

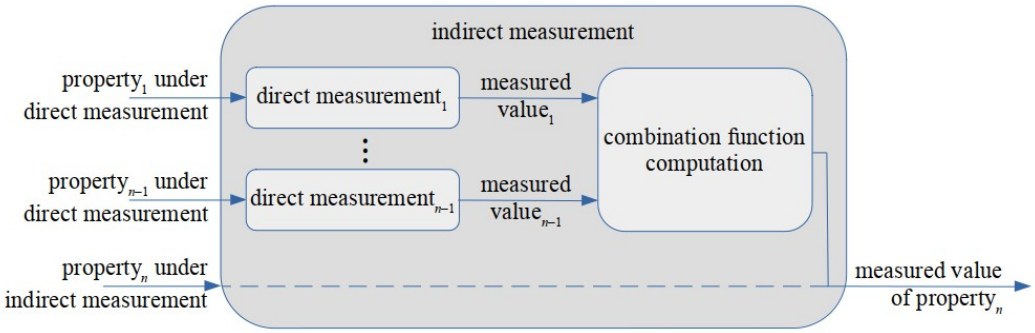


Fig. 7.2 The basic structure of an indirect measurement (measurement uncertainty is not considered here), including some unavoidable direct measurements (adapted from Fig. 2.11)

Even this simple analysis highlights the fundamental differences between direct and indirect methods of measurement, where a computation component is present in both, but with clearly different roles. Let us compare these methods in more detail in Table 7.1.

Table 7.1 A comparison of direct and indirect methods of measurement, with respect to the role of the computation component

In a *direct* (method of) measurement, the computation component	In an *indirect* (method of) measurement, which encapsulates one or more direct (methods) of measurement, the computation component
is a calibration function, f,	is a combination function, f,
which is a mathematical model *of the behavior of a measuring instrument*, with respect to the environmental properties that influence the relation between the property being measured and the instrument indication.	which is a mathematical model *of the measurand*, with respect to the way it is affected by, and more generally related to, other properties.
Such a model reconstructs the behavior of a measuring instrument, being the inverse of the instrument transduction function,	Such a model describes the relationship among the quantities of the object under consideration or of related objects, not the behavior of an instrument,
and in fact one of its arguments is the instrument indication.	and in fact its arguments do not include instrument indications.
The inverse of f, i.e., the transduction function, describes the cause-effect relationship realized by the instrument.	f does not necessarily involve cause-effect relationships.
The condition that the instrument needs to be calibrated corresponds to the fact that f is not completely known (for example, it could be parametric, and calibration itself generates the	Since f has nothing to do with instrument, it is known independently of the fact that there are instruments to be calibrated.

parameter values).	
Thanks to instrument calibration, from the value of instrument indications *f* computes a value for the measurand.	From the values of intermediate measurands, characteristic of the object under consideration or of related objects and not of an instrument, *f* computes a value for the measurand.

Of course, our point is not merely lexical: we emphasize the importance of maintaining a distinction between these two methods,[7] be they called "direct" and "indirect" or anything else.

7.2.2 Refining the distinction between direct and indirect measurement: first step

We have implicitly assumed so far that in direct measurements the property intended to be measured, i.e., the measurand (JCGM, 2012: 2.3), is the same as the property with which the measuring instrument interacts, which the VIM calls the "quantity being measured". However, as the VIM itself acknowledges (see Note 3 to def. 2.3), the measurand is not necessarily the same as the quantity being measured. On this matter we have introduced in Sect. 2.3 the distinction between

- the *intended property*: the measurand, i.e., the property that is intended to be measured and to which the measurement result is attributed, and
- the *effective property*: the property of the object under measurement with which the instrument actually interacts and that causes an effect through the transduction performed by the instrument.

This accounts for the ambiguity of the terms "measured property" and "property under measurement", which may refer to both the intended property and the effective property. This distinction is not maintained in daily measurements, in which the measurand is often not explicitly defined, because the interest is to measure "here and now" and nothing else, with the goal of obtaining information about the property that induces the transduction in the instrument.[8] But whenever measurement results are aimed at providing transferable information, such an implicit, indexical definition is no longer appropriate, and the model of the measurand needs to be improved, possibly by including in it the specification of values of properties by which it is affected. In Sect. 3.2 we have called such properties "affecting properties", because they are the properties that affect the measurand (it is unfortunate that the VIM does not have an entry, nor a term, for this concept). For example, the length of an iron rod is affected by temperature, because of thermal expansion, and therefore a measurand could be defined as the length of the rod at a given temperature, say 293.15 K.

It should be noted that, generally, in a given measurement context the affecting properties and the influence properties are not the same:

- the *affecting properties* are causally related to the measurand, and therefore are in principle independent of the measuring instrument (in the example above the volume of the rod is modeled in such a way that the temperature of the environment is an affecting property);

[7] The fact that in some cases the transduction is repeated and the measured value is computed as a statistic of the sample of indication values does not create a third method: from the structural point of view in which we are interested here, it remains unproblematically a case of direct measurement.

[8] Thus, for example, when measuring the weight of foods at the supermarket, the measurand is accepted to be the weight of the object put on the scale with no further specifications, though of course the measuring instrument is expected to be appropriately calibrated, including all relevant corrections.

- the *influence properties* are causally related to the transduction implemented in the measuring instrument, and alter its behavior in producing an indication in response to a given effective property

where, of course, a property may be both an affecting property and an influence property: this does not remove the principled distinction.

Since affecting properties enter the model of the measurand and influence properties enter the model of the instrument behavior, and since our provisional characterizations of direct measurement and indirect measurement are grounded on the distinction between these two models, as Table 7.1 highlights, our introduction of the distinction between affecting properties and influence properties requires us to refine the distinction between direct and indirect (methods of) measurement introduced above.

Let us then suppose that the measurand is defined as the volume of an iron body at the temperature of 293.15 K and that it has been established – by, say, an independent measurement – that the current temperature is 287.65 K, possibly with some measurement uncertainty, which is not relevant here. There are at least three possible strategies for coping with this situation:

S1: prior to measurement an empirical action is performed for changing the temperature of the body to the specified value, thus operating what could be called an *empirical correction*;[9]

S2: the difference between the specified and actual temperatures is taken into account by suitably *increasing definitional uncertainty*, for example by admitting that the measurand is defined within a given range of temperatures, and thus still independently of the way the measurement is performed; at the end measurement uncertainty is compared with definitional uncertainty to check that the former is not lower than the latter (which would mean that some resources were wasted in performing unnecessarily accurate measurements);[10]

S3: whether an empirical action is performed for changing the temperature of the body or not, both the current volume of the body and its current (thus possibly modified) temperature are measured, and then their values are properly combined, by a function that could be called a *computational correction*, to obtain a value for the measurand.

In all three of these strategies the measuring instrument is designed to empirically interact with properties of the same kind as the measurand, i.e., the volume of the body: according to the provisional characterization proposed above, these cases would then be all classified as direct measurements. However, their analysis in the light of the comparison proposed in Table 7.1 reveals some significant differences. Let us consider them with respect to one simple question: what if the model of the measurand (and more generally of the object under measurement) is wrong?

- S1 *does not* actually rely on the model of the measurand, as all that is required for justifying the choice of performing the empirical correction is the qualitative hypothesis that the measurand somehow depends on temperature; were the model discovered to be wrong – i.e., the volume of the body does not depend on its temperature as instead stated – this strategy would remain effective. When using this strategy, *measurement results do not depend on the validity of the model of the measurand*: even if volume did not depend on another property,

[9] Given the VIM definition of <adjustment of a measuring system> as a "set of operations carried out on a measuring system so that it provides prescribed [indication values] corresponding to given values of a quantity to be measured" (JCGM, 2012: 3.11, adapted), this action could be also called "adjustment of the measurement environment".

[10] As already mentioned in Footnote 13 of Chap. 3, definitional uncertainty is sometimes understood simply as one of the components of measurement uncertainty, and as such can be combined with the other components. We follow here the other approach, and consider it as the lower bound of the result of such a combination.

such as temperature as expected, the measurement would *directly* lead to a result, and the measurement uncertainty would not be affected by this;

- S2 *might* rely on the model of the measurand, but only for evaluating the contribution of the difference of temperatures to definitional uncertainty, what traditionally is called a "bias", i.e., the estimate of a systematic error (JCGM, 2012: 2.18) induced by such a difference; were the model discovered to be wrong, the only consequence would be that definitional uncertainty might be under- or over-evaluated. Under the assumption that definitional uncertainty is compared with measurement uncertainty, not propagated as a component of the uncertainty budget (see Sect. 3.2.4), also when using on this strategy *measurement results do not depend on the validity of the model of the measurand*;
- S3 *entirely* relies on the model of the measurand: measurement results are obtained by computing a combination function that models the relationship among the relevant properties of the object: were the relation between temperature and volume discovered to be significantly different from the one used to compute the correction, measurement results should be changed accordingly; moreover, the combination function is exploited for propagating the uncertainties of (effective) volume and temperature, to obtain the uncertainty of (modeled, i.e., intended) volume. When using this strategy, *measurement results do depend on the validity of the model of the measurand*.

Both S1 and S2 fulfill the conditions in the left column of Table 7.1 and therefore can be considered uncontroversial cases of direct measurement, thus showing that a measurement may be direct even if some affecting properties are acknowledged to be part of the model of the measurand. S3 may be described as the two (direct) measurements of (uncorrected) volume and temperature, followed by a computational correction for obtaining a value of (intended) volume by computing a combination function which is (part of) the model of the measurand. Hence this structure fits with the conditions in the right column of Table 7.1. However, S3 is not a typical case of indirect measurement, such as when the density of a body is measured by computing its value as a function of the values of the mass and the volume of the body. Hence this difference needs to be further analyzed.

7.2.3 Refining the distinction between direct and indirect measurement: second step

At the core of a model of a measurand is the hypothesis that the general property of which the measurand is an instance is an element of a network of properties with which the general property is in a lawlike relation (Sect. 6.6 presents a short analysis of this important subject). There are two basic reasons for exploiting such a network in calculating a value of the measurand, as exemplified by the two cases mentioned above:

- (example 1) the volume of an object is related to its temperature, i.e., the network includes volume and temperature; the definition of the measurand specifies a reference temperature, and the network allows us to correct the directly measured value of volume by taking into account the difference between the measured temperature and the specified temperature;
- (example 2) the density of an object is related to its mass and volume, i.e., the network includes density, mass, and volume, and allows us to compute a value of density as a function of the measured values of mass and volume.

Hence, both examples refer to the model of the measurand, but only the first has to do with the possible distinction between the effective property and the intended property, as generated by some affecting properties, given that we would surely not think of density as *affected* by mass and volume,

nor that mass and volume operate as corrections for density. This calls for a model-related refinement of the characterization of the distinction between direct and indirect (methods of) measurement proposed above. Let us then revise Table 7.1 accordingly by adding a middle column, where the left column is the same as in Table 7.1 and examples 1 and 2 are cases for the middle and the right columns respectively.

Table 7.2 A comparison of three general methods of measurement, provisionally called "method A", "method B", and "method C"

In a *method A* of measurement, the computation component	In a *method B* of measurement, the computation component	In a *method C* of measurement, the computation component
is a calibration function, f,	is a correction function, f,	is a combination function, f,
which is a mathematical model *of the behavior of a measuring instrument*, with respect to the environmental properties that influence the relation between the property being measured and the instrument indication.	which is a mathematical model *of the measurand*, with respect to the way the measurand is affected by other properties.	which is a mathematical model *of the measurand*, with respect to the way the measurand is related to other properties.
Such a model reconstructs the behavior of a measuring instrument, being the inverse of the instrument transduction function,	Such a model describes how the measurand depends on the effective property and the affecting properties, not the behavior of an instrument,	Such a model describes the relationship that the measurand has with other properties of the object under consideration or of related objects, not the behavior of an instrument,
and in fact the structure of f is $f(P_{ind}, ..., P_{infl\,i}, ...) = P_{eff} = P_{int}$, where P_{ind} is the instrument indication, $P_{infl\,i}$ is the ith influence property, P_{eff} is the effective property, and P_{int} is the intended property.	and in fact the structure of f is $f(P_{eff}, ..., P_{aff\,i}, ...)) = P_{int}$ where P_{eff} is the effective property, $P_{aff\,i}$ is the ith affecting property, and P_{int} is the intended property.	and in fact the structure of f is $f(..., P_{meas\,i}, ...)) = P_{int}$ where $P_{meas\,i}$ is the ith intermediate measurand and P_{int} is the intended property.
The inverse of f, i.e., the transduction function, describes the cause-effect relationship realized by the instrument.	f describes a cause-effect relationship between the effective property, the affecting properties, and the measurand.	f does not necessarily involve cause-effect relationships.

The fact that the instrument needs to be calibrated corresponds to the fact that *f* is not completely known (for example, it could be parametric, and calibration gives parameter values).	The fact that the value of the effective property needs to be corrected corresponds to the fact that the conditions in which the measurement is performed do not correspond to the conditions specified in the definition of the measurand.	Since *f* has nothing to do with instruments, it is known independently of the fact that there are instruments to be calibrated.
Thanks to instrument calibration, from the value of instrument indication, *f* computes a value for the measurand.	Thanks to correction, from the measured value of the effective property, *f* computes a value for the measurand.	From the values of intermediate measurands, characteristic of the object under consideration or of related objects and not of an instrument, *f* computes a value for the measurand.

It should be noted that all three methods include a mathematical model in the form of a function by means of which a value for the measurand can be calculated and its uncertainty evaluated, with the consequence that measurement cannot be a purely empirical process: none of these methods can provide "pure data". Only in an abstract perspective, however, methods A, B, and C may be treated in an undifferentiated way, under the consideration that all of them require calculating a function "among all properties known to be involved in a measurement", paraphrasing from the previously-quoted VIM definition of 'measurement model' (JCGM, 2012: 2.48). Although the same formal rules for, say, uncertainty propagation apply to the three cases, maintaining the distinctions presented in Table 7.2 seems to be helpful for a better understanding of the structure of the measurement process and the role of mathematical models in it.

This analysis shows that measurement methods can be classified according to two general criteria, related

- (first criterion) to the way in which the measuring instrument is designed and coupled with the object that carries the measurand, and
- (second criterion) to the way in which the measurand is modeled and this model is exploited to compute the measurement result.

In light of this distinction and in reference to the content of Table 7.2 again, *method A (left column) and method C (right column) may be acknowledged to be direct and indirect respectively*, according to these characterizations, where parts (*i*) and (*ii*) in the two definitions below are based on the first and the second criterion respectively:

> *a measurement is based on a direct method (as in the left column* of Table 7.2*) when*
> *(i) an instrument is used that is coupled with the object carrying the measurand and is designed to interact with instances of the general property of the measurand, and*
> *(ii) the model of the measurand is only used in measurement for identifying the measurand.*

and:

a measurement is based on an indirect method (as in the right column of Table 7.2) when
(i) instruments are used that are not necessarily coupled with objects that carry the measurand or designed to interact with instances of the general property of the measurand, and
(ii) the model of the measurand is used in measurement for identifying the measurand and its dependence on the properties from which the measurement result can be computed.

However, method B (middle column of Table 7.2) is not uniquely characterized, being analogous to a direct method with respect to the first criterion and to an indirect method with respect to the second criterion. Since we are not interested here in lexical issues, we simply call this method "direct/indirect":

a measurement is based on a direct/indirect method (as in the middle column of Table 7.2) when
(i) an instrument is used that is coupled with the object carrying the measurand and is designed to interact with instances of the general property of the measurand, and
(ii) the model of the measurand is used in measurement for identifying the measurand and its dependence on affecting properties, from which the measurement result can be computed.

This mixed case highlights the complexity of our subject, and perhaps explains some of the confusion around the distinction between direct and indirect methods of measurement. Furthermore, the claims that any measurement is based on either a direct or an indirect method, in the sense proposed here, and that any indirect measurement requires at least one direct measurement, contribute to the clarification of the strategic issue of whether measurement science is becoming a branch of data science: the answer is negative. Even though the computational components are becoming more and more important, measurement maintains its distinction from computation by virtue of involving empirical components: *any direct measurement includes at least one empirical component, and any measurement includes at least one direct measurement*.

Hence, the framework we are going to propose needs to embed a structural model of direct measurement and to be based on it.

7.3. A structural model of direct measurement

We are grounding our account on the model proposed by Giordani and Mari (2019), which shares important features with the Berkeley Assessment System (BAS) model (Wilson, 2005; Wilson & Sloane, 2001), and Evidence Centered Design (ECD: Mislevy et al., 2003). We initially use the same example as in Giordani and Mari to help make the description of the structural model clearer and more concrete. The aim of this section is to supply a structural model for the fulfillment of the Basic Evaluation Equation for a generic property, in the case that in the previous section we have characterized as *direct measurement*.

In reference to what we have presented in Sect. 2.3, our starting point is the consideration that measurement is a process that maps empirical entities to informational entities, i.e., empirical properties of objects to values, and this highlights the fundamental role that *scales* have in measurement (and more generally evaluation) processes, as also introduced in Box 6.2. What could be called a Basic Evaluation Scale[11] is then a mapping

[11] The adjective "basic" refers to the simplification of not taking uncertainty into account in scale construction.

property of a given object → identifier of a property

from a set of comparable and distinguishable properties of objects, all of them being instances of the same general property, to a set of distinct identifiers each corresponding to a value of that general property, where the mapping is constrained by the conditions of scale transformation (see Sects. 6.3.2 and 6.5.1). For example, for ordinal evaluations scales are defined up to monotonic transformations among identifiers, and for ratio evaluations scales are defined up to changes of the units. While syntactically equivalent to a Basic Evaluation Equation, a Basic Evaluation Scale differs in that it is not, in principle, true or false, depending on which identifier, and then which value, is chosen for each given property. Rather, it is a specification, which is instead required to be consistent: that is, if the property $P[a]$ is greater than the property $P[b]$ then the value specified for $P[a]$ must be greater than the value specified for $P[b]$, and so on. This can be summarized as in Table 7.3.

Table 7.3 A comparison of Basic Evaluation Equations and Basic Evaluation Scales

A Basic Evaluation Equation	A Basic Evaluation Scale
is related to the property of a single object,	is related to a set of comparable and distinguishable properties of objects,
is in principle true or false,	is in principle consistent or inconsistent,
and is the simplest case of a measurement result.	and is the simplest case of the outcome of a scale construction.

Two scales are involved at the core of any direct measurement. In the example of the measurement of the temperature Θ of an object a, $\Theta[a]$, by means of an alcohol thermometer, we can consider the following.

- One is the scale that maps the temperatures of some already-established measurement standards to their values. For temperature, there would be a set of standards of temperature $\{s_j^*\}$, each having a temperature $\Theta[s_j^*]$, and a Basic Evaluation Scale of temperature would be built upon it:[12]

  for each s_j^* in the given set, $\Theta[s_j^*] \rightarrow \theta_j$

  where θ_j is the j-th value in the scale.[13] For example, bodies at the boiling and freezing points of water, at sea-level, could be two such temperature standards, and the numbers 100 and 0 in the given scale could be the chosen values. Since measurement standards are expected to be socially available for supporting metrological traceability (see Sect. 3.3.1), a scale about measurement standards may be called a *public scale*.
- The other is the scale that maps the instrument indications for a specific instrument (or type of instrument) to indication values. For an alcohol thermometer, this would map the positions of the upper surface of the alcohol in the tube of the thermometer to position values: thus, a set of

[12] For the sake of simplicity, the distinction between scale identifiers and values is not maintained here, so that for example the identifier 20 in the Celsius scale and the value 20 °C are treated as being as substantially the same entity. See Box 6.2 for a further analysis of this.

[13] In the human sciences readily transportable measurement standards are not very common. In some contexts, synthetic reference objects have been found to be useful, such as computerized chess players (Maul et al., 2019).

positions X_i^* of the upper surface of the alcohol in the tube of the thermometer is chosen, and a Basic Evaluation Scale of position is built upon it:

for all X_i^* in the given set, $X_i^* \rightarrow x_i$

where x_i is the i-th value in the scale (in millimetres, say). This scale is specific to the given measuring instrument, and therefore may be called a *local scale*.[14]

The fundamental structure of a direct measurement is *in the relation between a public scale and a local scale*:

- the measurand is a property of the same kind as the properties in the public scale, and measurement needs to produce measured values in the public scale;
- the measuring instrument performs a transduction from the property under measurement to a property of the same kind as the properties in the local scale (e.g., for an alcohol thermometer, from temperature to positions), so that indication values are values in the local scale (say, millimetres).

As we will see, the calibration of a measuring instrument may be interpreted as the process of connecting a public scale with the local scale of the instrument. It is this connection that makes it possible for a calibrated instrument to obtain a measurand value from an instrument indication value, i.e., a value in a public scale from a value in a local scale, as depicted in Fig. 7.3.

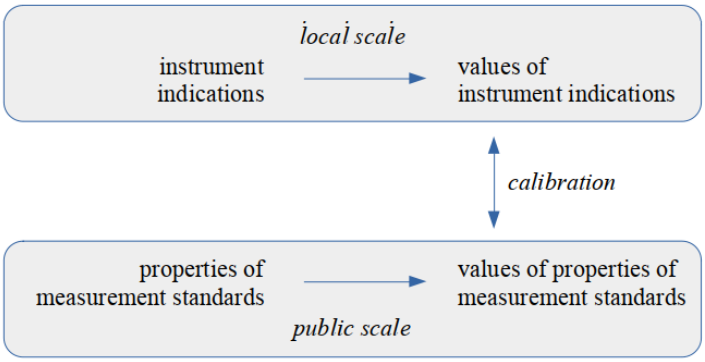


Fig. 7.3 Calibration as a relation of a public scale and a local scale

It is upon this idea that our structural model of direct measurement is grounded. We next describe its components and how they are connected with each other, first in the simpler case in which uncertainties are not included in the description and starting from the preliminary process of the design and construction of the measuring instrument.

7.3.1 The design and construction of a measuring instrument: backgrounder

Direct measurement is enabled by the use of a measuring instrument, a device able to interact with the property under measurement[15] and to map it to a value in the local scale embedded in the

[14] By presenting public scales first and then local scales we are following a conceptual sequence from what is outside the box, i.e., a measuring instrument, to what is inside the box. But in the historical development public scales may be the outcome of the previous development of multiple local scales and the assessment of their agreement. In the example of temperature, the first public thermometric scales (e.g., Celsius and Fahrenheit) were developed only after several thermometers were discovered to be in substantial agreement in their behavior (Chang, 2007).

[15] For the sake of simplicity, in this initial presentation we do not distinguish between (i) the intended property (that is, the measurand), i.e., the property referred to in the Basic Evaluation Equation that reports the result of measurement, and (ii) the effective property, i.e., the property that interacts with the measuring instrument and produces an effect on it; we call both of them the *property under measurement*. The distinction is recovered and introduced in the model in Sect. 7.4.2 in the context of the analysis of the role of definitional uncertainty.

instrument. The usual structure of a measuring instrument may be described as constituted of three functional components: (i) a transducer, (ii) a scale, and (iii) something that matches the transducer outputs with the properties in the scale (an exception is discussed in Sect. 7.3.3).

The design and construction of a measuring instrument are grounded on the consideration of empirical properties of objects as associated with modes of empirical interaction of the objects with their environment, as introduced in Sect. 2.2.3 and then commented in Sect 5.1, and therefore on the formulation of the hypothesis – corroborated by appropriate observations – of a causal relationship between the general property of interest and another general property – the transduced property – whose instances are in some sense more readily empirically distinguishable as state transitions of the instrument. As described by Robert Rosen, "every recognition, measurement, discrimination or classification ultimately rests on the capacity of a given system S to induce a dynamics (i.e., a change of state) in another system M; this latter system will be variously called a meter, discriminator, recognizer, classifier, etc. It is this dynamical behavior in M, and particularly its asymptotic behavior, that provides the basis for learning about and describing the original system S." (1978: p. x).

In the case of temperature this happened with the discovery of thermal expansion, i.e., the transduction effect according to which changes of the temperature of a body cause changes in its volume. In the case of a competence, like reading comprehension ability (RCA), this is typically based on the construction of a test whose items are specifically designed for checking that competence, where the transduced property is then the pattern of responses produced by a reader who responds to those items. We develop the case of temperature here, and a psychosocial case in Sect. 7.3.5.

Let us consider the example of an alcohol thermometer: it is a transducer from temperatures Θ of objects a, $\Theta[a]$, to positions X_m of the upper surface of the alcohol column housed in the glass tube of the thermometer, where the index m refers to the measuring instrument. Under a hypothesis of causality, the transduction is modeled as an empirical map $\Theta[a] \rightarrow X_m$. Prior to its usage in a measurement, the instrument must then be configured by etching a set of marks along the tube, corresponding to the distinguishable positions X_i^* of the upper surface of the alcohol in the tube, in such a way that these positions – which could be called *local reference properties* to emphasize their dependence on the instrument – are mapped to values, so as to establish a local scale, modeled as an empirical-to-informational map $X_i^* \rightarrow x_i$ from reference positions X_i^* to values of position x_i. This map, which is constructed under controlled conditions and then is assumed to be invertible, takes into account the specific features of the instrument, and as such differs from instrument to instrument.

A transducer and a local scale are the main functional components of a measuring instrument. For coupling them, the output of the former needs to be matched with the input of the latter. In an alcohol thermometer this requires being able to establish the reference position X_i^* that best matches the transduced position X_m, an operation modeled by a map $X_m \rightarrow X_i^*$.

On this basis the stages of the process of direct measurement are designed and performed, including the construction and the application of the instrument calibration function, an informational map $x_i \rightarrow \theta_j$ from instrument-related, and therefore local, values of position x_i to the sought values of temperature θ_j.

7.3.2 The stages of direct measurement

Transduction. The first empirical operation of a measuring instrument is a transduction, in our example from temperatures to positions of the upper surface of the alcohol in the thermometer tube, where this mapping is based on an *assumption of causality*: that is, via the transducer, the transduced

property is the effect of the property under measurement.[16] This behavior is the outcome of the two basic features of a measuring instrument: it is sensitive to a given property, and it can change its state as the result of its interaction with an object carrying that property. From this perspective, a measuring instrument is a device designed so as to empirically realize a cause-effect relationship as an input-output function, the input being a property of external objects with which the instrument is able to interact and the output being a property of the instrument itself. Thus, suppose we bring the thermometer m into contact with the object a; then, the position of the upper surface of the alcohol is X_m, which is the transduced property obtained by the transduction of $\Theta[a]$ via the map $\Theta[a] \rightarrow X_m$, as depicted in Fig. 7.4.

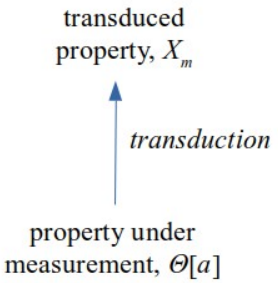


Fig. 7.4 The first stage of direct measurement: transduction (in the example $\Theta[a]$ is the temperature of an object and X_m is a position of the upper surface of alcohol in the tube of a thermometer)

Matching. The output of a transducer is an empirical property. Transduced properties are not always unambiguously distinguishable with each other. However, measuring instruments are designed so as to make the identification of transduced properties possible, and usually effective, through their matching with a predefined set of reference properties X_i^*. This is traditionally performed by the measurer, who visually compares the transduced property and some references properties somehow identified on the instrument – for example the position of the upper surface of alcohol in the tube as matched to the marks etched along the tube – and to the best of her ability tries to minimize indication errors, such as those due to insufficient lighting, parallax, etc. In electronic instrumentation this task is performed by a quantizer (usually a component of an analog-to-digital converter, together with a sampler), which operates as a classifier of the transduced property. Whether manual or automatic, the matching can be then modeled as a map $X_m \rightarrow X_i^*$, as depicted in Fig. 7.5. In the case of transducers designed to produce already discrete properties, the matching function might be, trivially, the identity.

[16] This assumption of causality is at the core of the very possibility of measurement by means of these kinds of devices, which are instances of what Nancy Cartwright calls a *nomological machine*, "a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behaviour that we represent in our scientific laws." (1999: p. 50).

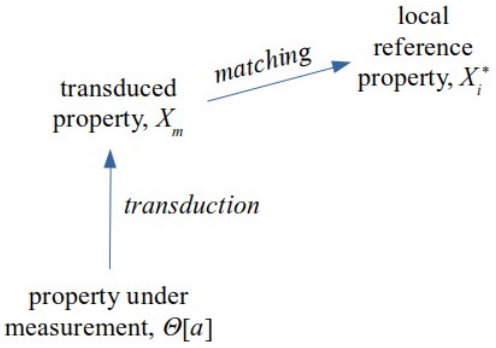


Fig. 7.5 The first two stages of direct measurement: after transduction, matching (in the example X_m is a position of the upper surface of alcohol in the thermometer tube and X_i^* is the position of the i-th mark etched on the scale of the thermometer)

Local scale application. Though still empirical, local reference properties can be effectively dealt with thanks to the appropriate design of the measuring instrument, as it is in the case of the positions identified by etches along the tube of a thermometer. In particular, this allows the instrument manufacturer or its users to establish an instrument-specific, and hence local, scale – i.e., an empirical-informational map from local reference properties to identifiers – and then values of the transduced property (see the analysis on evaluation scales in Box 6.2). Hence, once the reference property X_i^* is identified that best matches X_m, the map $X_i^* \rightarrow x_i$ is applied to find the local value that, via the instrument, corresponds to the property under measurement $\Theta[a]$, as depicted in Fig. 7.6.

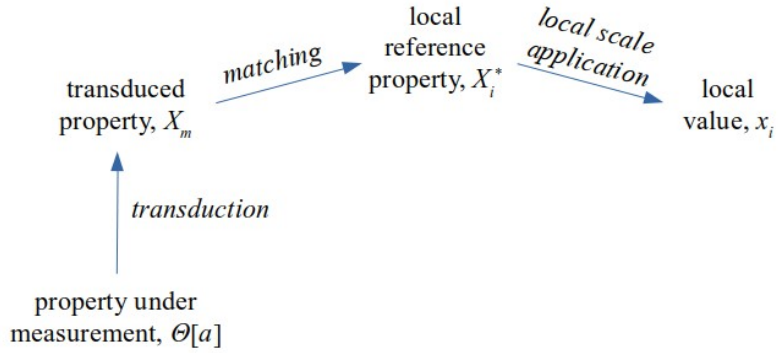


Fig. 7.6 The first three stages of direct measurement: after transduction and matching, local scale application (in the example X_i^* is the position of the i-th mark etched on the scale of the thermometer and x_i is a value of position)

As Fig. 7.7 shows, a sequence of actions

transduction → matching → local scale application

allows us to attribute local, i.e., instrument-specific, values to properties under measurement. While this is already a map from empirical to informational entities, as expected from a measurement, there are two main drawbacks in this:

- the values are in the instrument scale, not the measured property scale: in our example, this would imply reporting information about temperature in terms of values of length;
- the values depend on the specific instrument, not only the property under measurement: in our example, this would imply producing information that is valid only for a given thermometer.

Such a "local measurement" is called *pre-measurement* by Frigerio et al. (2010).[17]

[17] This is sometimes referred to as the "pilot" stage of instrument construction in the human sciences.

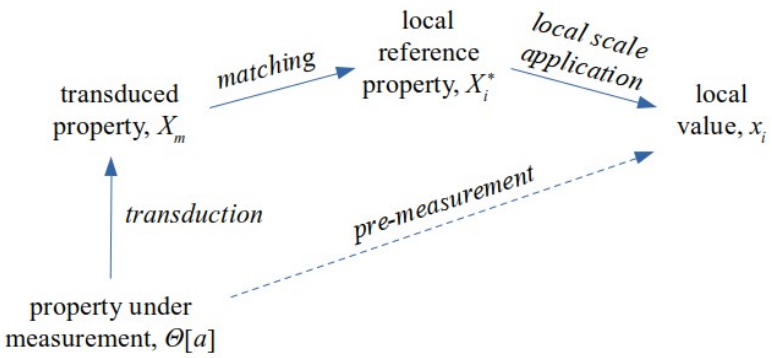


Fig. 7.7 Pre-measurement, as the composition of transduction, matching, and local scale application

The drawbacks of pre-measurement can be overcome by introducing a public scale for the property under measurement and calibrating the instrument against it.

Public scale construction. As discussed above, a set of reference properties of the same kind as the property under measurement is made available through a set of measurement standards, and a public scale is thus built as a map from such reference properties to identifiers or values (see the analysis on evaluation scales in Box 6.2). In the case of temperature, $\Theta[s_j^*] \rightarrow \theta_j$ is a public scale that maps each reference temperature $\Theta[s_j^*]$ of the standard s_j^* to the value θ_j, as depicted in Fig. 7.8. A traditional way to accomplish this is by means of two standards, such that $\Theta[s_1^*]$ is the temperature of the freezing point of water and $\Theta[s_2^*]$ is the temperature of the boiling point of water, under given conditions of pressure, to which the values 0 and 100 are conventionally assigned in the Celsius (Centigrade) scale, where the values to be associated to any other temperature can be obtained by means of some sort of interpolation, for example as discussed in Sect. 6.3.6.

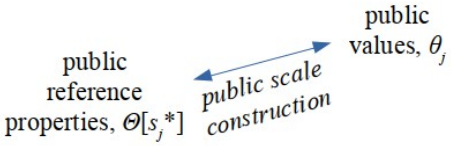


Fig. 7.8 A preliminary stage of direct measurement: public scale construction (in the example $\Theta[s_j^*]$ is the temperature of the measurement standard s_j^* and θ_j is a value of temperature)

Calibration. The sequence

transduction → matching → scale application

can be applied not only to a property under measurement, but also to the reference properties of a public scale. What is obtained are still local values, but in this case they are known to correspond to the public values that were attributed to the reference properties. A pointwise mapping from local values to public values can be then established by repeating the process for each reference property: this is the instrument *calibration map*. In our temperature example, the chain $\Theta[s_j^*] \rightarrow X_m \rightarrow X_i^* \rightarrow x_i$ together with the scale $\Theta[s_j^*] \rightarrow \theta_j$ lead to the calibration map $x_i \leftrightarrow \theta_j$, as depicted in Fig. 7.9.[18] Such a

[18] Under the hypothesis that the instrument behavior is modeled by an analytical function, and the underlying structure of the evaluation is sufficiently rich (see Sect. 6.5.3), this pointwise map may be interpolated, so as to obtain pairs (local value, public value) also for public values for which the instrument was not directly calibrated. For example the thermometer could be hypothesized to behave in a linear way between the freezing and the boiling point of water – i.e., a change of the measured temperature produces a proportional change of the position of the upper surface of alcohol in the tube – so that the midpoint of the upper surface would be mapped to 50 °C, and so on. Of course, the thermometer could be calibrated even if its behavior were not linear, but this would require the availability of other public reference

map may be presented by means of a table, listing the relevant pairs (x_i, θ_j), or a (local values, public values) chart, through which any obtained indication value can be associated with a measured value. In some cases, the preferred option is instead to embed the information produced by the calibration directly in the instrument, for example by writing the public values directly on the instrument scale in place of the indication values, as is common for many thermometers. Of course, this substitution does not imply any structural change in the model we have just presented.

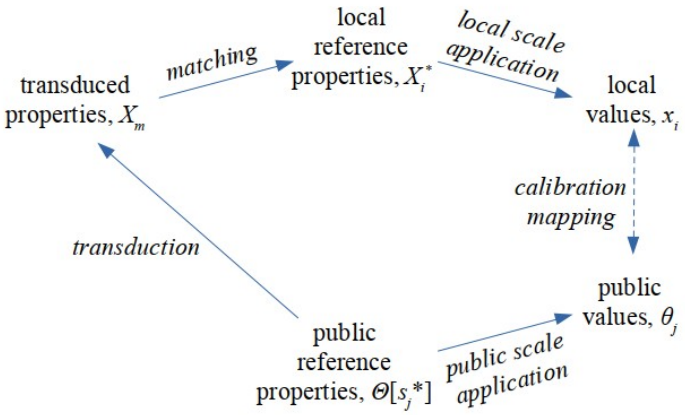


Fig. 7.9 A preliminary stage of direct measurement: calibration

Thanks to the calibration map, pre-measurement finally upgrades to (direct) measurement: the instrument-specific map of pre-measurement $\Theta[a] \rightarrow x_i$ can be composed with the calibration map $x_i \leftrightarrow \theta_j$, together producing a map $\Theta[a] \rightarrow \theta_j$, which corresponds to a Basic Evaluation Equation and therefore to the expected outcome of a measurement (again, in the simple case in which uncertainties are not taken into account). This structure is depicted in Fig. 7.10

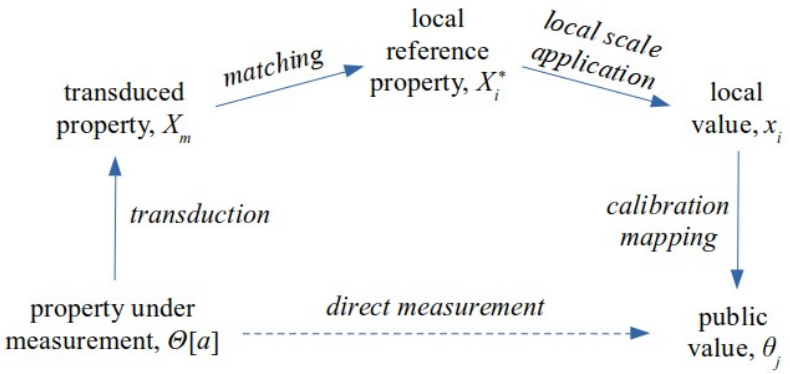


Fig. 7.10 The operational structure of direct measurement, as a composition of pre-measurement and calibration

which is a summary version of the more explicit structure in Fig. 7.11, in which the role of the public scale for the creation of the calibration map is also shown.

temperatures, intermediate between the freezing and the boiling point of water.

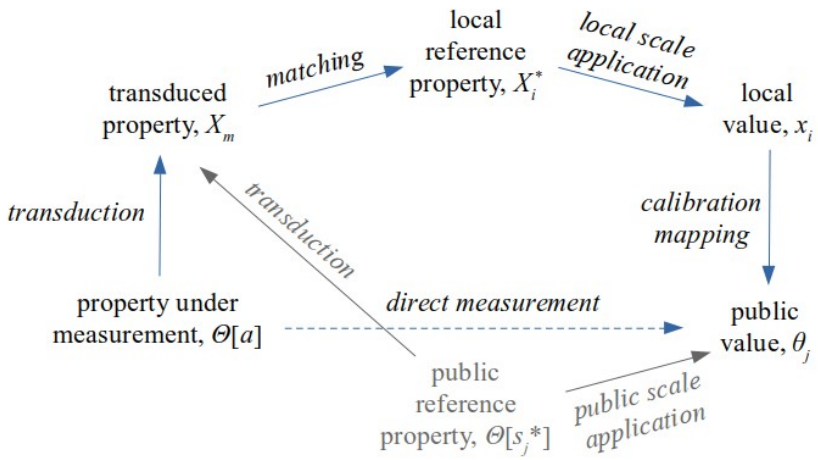


Fig. 7.11 The operational structure of direct measurement, highlighting the role of the public scale for the creation of the calibration map; note that the two transductions are performed at different times

7.3.3 An alternative implementation

As the diagram in Fig. 7.11 shows, in measurements like the ones performed by means of an alcohol thermometer the properties under measurement and the reference properties in the public scale *are not compared directly*: as a consequence, for the instrument map and the calibration map to be correctly composed, the instrument must not significantly change its behavior over time or across environments. These two conditions were characterized, in Sect. 3.2.1, in terms of instrument stability and selectivity, respectively, such that if these conditions do not hold the instrument may be in need of recalibration or repair. Under the condition that the instrument behaves in a sufficiently stable and selective way, the sequence shown in Fig. 7.10 of

transduction → matching → local scale application → calibration mapping

provides an operational implementation of a direct measurement.

An alternative implementation of direct measurement is possible if the instrument allows the *direct comparison of* the property under measurement and the public reference properties, as in the case of the measurement of weight by means of a two-pan balance,[19] in which the property under measurement, i.e., the weight of the object under measurement, is directly compared with some standard masses.[20] This process may be modeled as matching the property under measurement to the reference properties in the public scale. The operational structure of direct measurement in this case is much simpler, as depicted in Fig. 7.12.

[19] We are using the contrast of 'direct' and 'indirect' in two distinct ways. Given the distinction between direct and indirect measurement, the reference is here to a difference about operational implementations of direct measurement, which may be performed by indirectly or directly comparing the measured property and the reference properties in the public scale. In summary, a direct measurement can be performed through an indirect (as usual) or a direct comparison.

[20] This case is quite rare in the human sciences, due to the already mentioned rarity of reference objects.

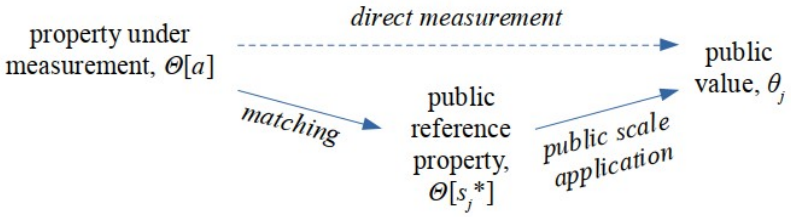


Fig. 7.12 The operational structure of direct measurement, as based on the direct comparison of the property under measurement and the public reference properties

7.3.4 The Hexagon Framework

The two operational implementations presented in the previous sections can be merged into one conceptual structure, which then presents a more implementation-neutral and therefore general interpretation of direct measurement, as depicted in Fig. 7.13, referred to henceforth in this book as the "Hexagon Framework". This structure shows that in a direct measurement the map

property under measurement → public value

may be implemented, and therefore a Basic Evaluation Equation may be obtained, in two alternative ways:

transduction → matching → local scale application → calibration mapping

and

matching → public scale application

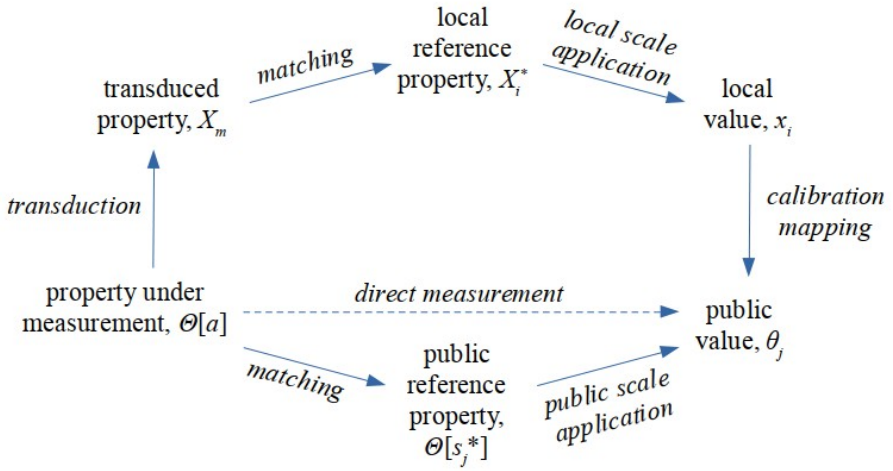


Fig. 7.13 The conceptual structure of direct measurement: the Hexagon Framework

This diagram reveals some symmetries in the conceptual structure of direct measurement according to the model we are proposing:

- there are a local scale and a public scale, connected by calibration (Fig. 7.14);[21]

[21] As noted above, and illustrated in Sect. 7.3.3 and in particular in Fig. 7.12, significant parts of this figure and the next two are absent in the case of direct comparisons with reference objects.

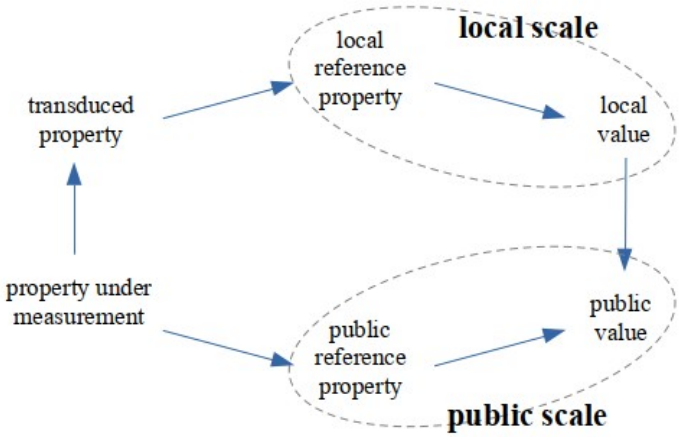


Fig. 7.14 First symmetry in the Hexagon Framework: local scale and public scale

- the transduction map has an informational counterpart in the calibration map (Fig. 7.15):

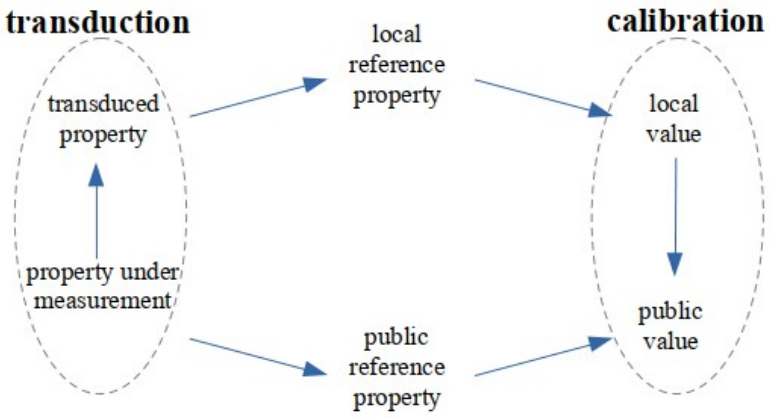


Fig. 7.15 Second symmetry in the Hexagon Framework: transduction and calibration

- measurement includes an empirical component and an informational component (Fig. 7.16):

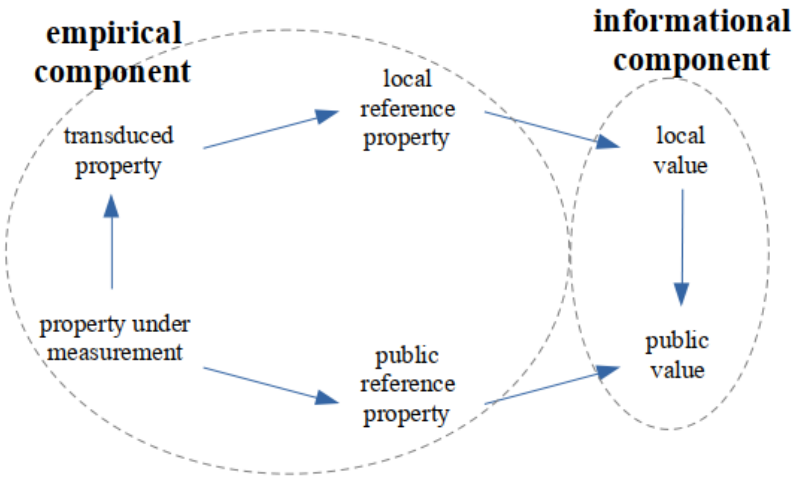


Fig. 7.16 Third symmetry in the Hexagon Framework: empirical component and informational component

This structural characterization of the process of measurement is independent of the algebraic structure of the scale. As discussed in Sect. 4.4, the model characterizes direct measurement through empirical but not mathematical conditions, and as such this structural account applies also to non-quantitative properties (see also Mari et al., 2017).[22] Furthermore, the fact that the characterization is purely structural makes the model applicable to understanding the measurement of both physical and non-physical properties. Let us see an example.

7.3.5 An example application of the model in the human sciences

In this section, we give an account of two aspects of the application of the Hexagon Framework, in the human sciences. First, we describe how the model can be applied to a recent example from educational measurement. Second, we then discuss how the perspective of the Hexagon Framework can be exploited to inform an instrument development process used on psychosocial measurements using the same example.

The common background context for both of these is an assessment system built for a middle school statistics curriculum that leans heavily on the application of learning sciences ideas in the STEM (science, technology, engineering, mathematics) domain, called the *Data Modeling* curriculum (Lehrer et al., 2014). The *Data Modeling* project[23] created a series of curriculum units based on real-world contexts that would be familiar and interesting to students, with the goal of making data modeling and statistical reasoning accessible to students who usually do not do so well on this subject. Within specific topic areas, the curriculum and instructional practices use the idea of a *construct map*, as described below, to help materials developers design and develop the educational materials, to help teachers guide the design of their instructional plans in the topic to be taught, and, where the students are sufficiently mature, to help learners appreciate their own growth in knowledge and skills. The construct map presented in this example describes transitions in reasoning about data modeling and statistical reasoning when middle school students are inducted into practices of visualizing (i.e., they draw illustrations of what they think is going on), carrying out scientific practices (e.g., they pay attention to the process, being careful to record what they observe), and modeling the variations they have observed and recorded in the context they are studying (e.g., they observe how different students' measurements of certain hard-to-measure properties of certain objects will vary, such as the heights of tall trees). In the *Data Modeling* curriculum, teaching and learning are closely coordinated with assessment, which is the focus here.

To make the idea of a construct map concrete for the reader, let us consider an instrument aimed at measuring student knowledge and skills – we call them a "competence" – in a part of the *Data Modeling* curriculum called *Modeling of Variability*, to be described in detail in the next paragraph. We assume that this student's property can vary at least ordinally, from high competence to low competence, and that the measurement developer can postulate a (finite) number of consecutive educationally distinguishable points between the extremes. This corresponds to a typical step in the development of the measurement of a property in the human sciences, where, before one has achieved the ability to carry out a full measurement process, as in the Hexagon Framework, the measurement developer has to go through a process of deriving an ordinal understanding of the property – in this case a competence of individuals – using the research literature and professional knowledge in the

[22] There is in fact a mathematical condition here, i.e., that the mappings involved in the process can be formalized as functions, but this does not constrain the measurability of non-quantitative properties. As discussed further in Sect. 7.4, this is to be reconsidered if uncertainties are included in the model.

[23] The project was carried out jointly by researchers at Vanderbilt University and the University of California, Berkeley, and was funded by the US National Science Foundation (NSF).

topic area. Quite often the property will be conceptualized as describing successive points in a typical process of *change*, or *development*, over time within an individual student, and the construct map can then be thought of as being analogous to a qualitative "roadmap" of change in the competence (see for example Black et al., 2011). In recognition of this analogy, these qualitatively different locations along the construct are called "waypoints": it is from these qualitative descriptions that the ordering of the waypoints is derived, and is very important to the instrument development process and also in the interpretation of the measurement results. Each waypoint has a qualitative description in its own right, but, in addition, it derives meaning by reference to the waypoints below it and above it. Finally, we model this competence as a dense property, in the sense that, in principle, for any two students with distinguishable competences there can be a third student whose competence is situated between theirs.

The property under measurement. The property that we focus on for this example is the competence to deal with Models of Variability, "MoV" for short. This is a student competence at a relatively young age, between about kindergarten and Grade 5. The project's final construct map for MoV is illustrated in Fig. 7.17, where the waypoints are symbolized by blue dots. At the low end of student development (i.e., at the bottom of the figure), the focus is on the identification of sources of variability, then advancing to the incorporation of devices to represent the mechanisms of those sources of variability; at the highest waypoint, students are also able to develop models of variability and to judge how well they work by examining how repeated model simulations relate to an empirical sample. Note that one waypoint has two labels attached to it (MoV2 and MoV3): this means that these are two different categories of student thinking that occur at the same point, or at least very similar points, in development (more about this later).

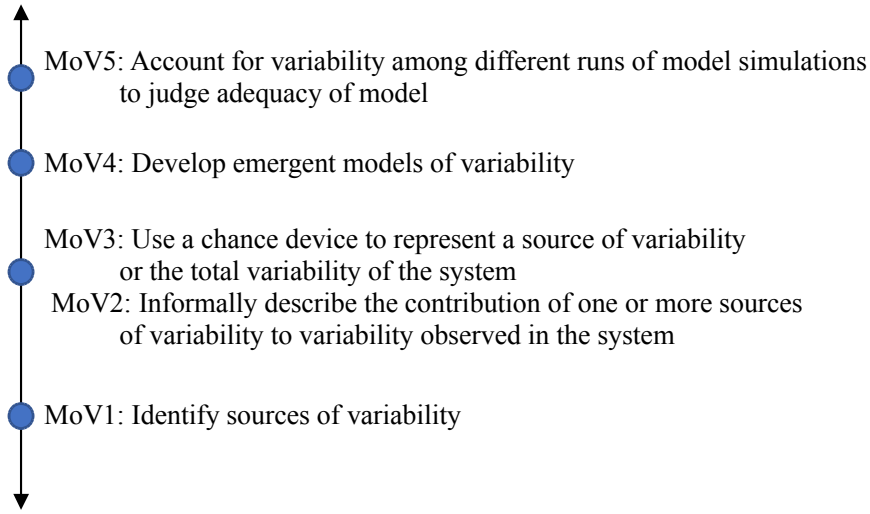
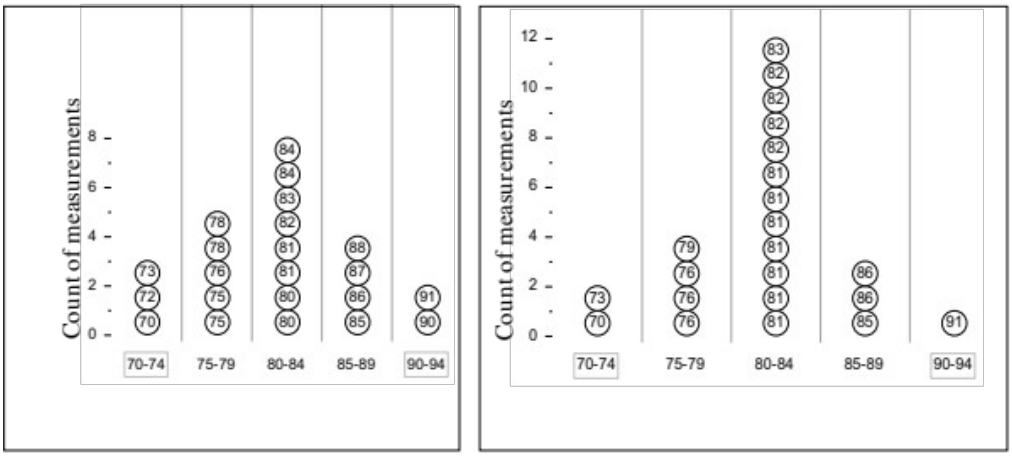


Fig. 7.17 The MoV construct map (the distances shown between waypoints are arbitrary, and they are not represented as equally far apart to emphasize that there is not assumption of "equal differences" in this diagram)

A group of musicians measured the width of a piano in centimeters. Each musician in this group measured using a small ruler (15 cm long). They had to flip the ruler over and over across the width of the piano to find the total number of centimeters. A second group of musicians also measured the piano's width using a meter stick instead. They simply laid the stick on the piano and read the width. The graphs below display the groups' measurements.



Piano width (cm) using the **small ruler** Piano width (cm) using the **meter stick**

1(a) The two groups used different tools.
Did the tool they used affect their measurement? *(check one)*

( ) Yes ( ) No

1(b) **Explain your answer.**
You can write on the displays if that will help you to explain better.

2(a) How does using a different tool change the precision of measurements? *(check one)*

(a) Using different tools does **not** affect the precision of measurements.
(b) Using the small ruler makes precision better.
(c) Using the meter stick makes precision better.

2(b) **Explain your answer.**
(What about the displays makes you think so?)

Fig. 7.18 The *Piano Width* task

Transduction. A task like the one shown in Fig. 7.18, as based on the MoV construct map, is designed to operate as a transducer, that is sensitive to the student's MoV competence[24] and produces a complex transduced property in the form of responses given to questions by each student, for example a response identifying an affirmative or a negative response for Question 1(a) and a written text for Question 1(b). This can be then modeled as a map from the property under measurement $\Theta[a]$ to a

[24] Of course, the task as such is not able to perform any transduction, which has to be driven by the student who provides her responses to the given questions. However, this is typical of all passive instruments, which are activated by the object under measurement or an external source. Also in this regard a MoV task is analogous to an alcohol thermometer, in which the upper surface of alcohol in the tube changes its position only because of the energy transferred from the object whose temperature is under measurement.

transduced property X_m, where $\Theta[a]$ is the MoV competence Θ of student a, as in Fig. 7.4. This task capitalizes on *Data Modeling* students' experiences in learning about measuring properties of awkward objects as a process that generates variability. In particular, Question 1(a) is intended to prompt the student take one of two positions. Some students may note that, for the given bins, the mode is the same for both displays and consider that sufficient to say "No": these students are not able to specify the source of the variation, as they are not perceiving the spread as being relevant. Following this set-up, the later Questions, 1(b) and 2(b), explore students' understanding of how measurement techniques could affect variation, providing the opportunity to gather evidence regarding whether a student's MoV competence is at least at the waypoint MoV2 in the construct map. This task does not provide opportunities for evidence above MoV2, as no models of chance or chance devices are involved in the question: hence, the sensitivity of this measuring instrument to changes of MoV competence above MoV2 is zero.

Matching. Particularly if the task is presented to students on paper, so that they operate writing by hand, the response produced by each student must be matched to the local (i.e., instrument-related) reference properties X_i^* for the item, and this can be modeled as a map $X_m \rightarrow X_i^*$, as in Fig. 7.5. In the simplest cases, as for Question 1(a), the transducer is designed to discriminate between two properties, X_1^* and X_2^*, corresponding to two alternative states of selection, as shown in Fig. 7.19. The matching may be performed by a human rater or by a mark sense reader. In practice, it may be that some responses are not uniquely classifiable by the local references, such as when a reader makes a mark that is not clearly in one box rather than another, which might lead to the addition of a third reference property, corresponding to such an "unclear" situation. Furthermore, for some open-ended questions the very possibility to define a set of reference properties to classify responses could be problematic: in these cases the process must include a well-documented and monitored system for judging the open-ended responses (see, for example, Wilson, in press: chapters 4 and 7) in order to fulfill the minimal conditions this Framework sets to consider it a measurement.

(X) Yes ( ) No ( ) Yes (X) No

Fig. 7.19 The local references for the example: X_1^* (left) and X_2^* (right)

Local scale construction and application. The local reference properties are still empirical entities, being for example patterns which may be read on a sheet of paper. However, the fact that they are preliminarily identified and guaranteed to be distinguishable in the instrument operation allows us to associate them with information entities. A map $X_i^* \rightarrow x_i$ from local reference properties X_i^* to local values x_i is then defined, as in Fig. 7.6. Typically, for each dichotomous item the patterns corresponding to the correct response are scored as 1 and the patterns corresponding to the incorrect response are scored as 0. Once such a local scale has been constructed, whenever one of its reference properties X_i^* is indicated by a student's response, the corresponding value x_i is recognized by applying the scale. As it is commonly said, the student's response is "scored", and this is the conclusion of the pre-measurement, as depicted in Fig. 7.7. For more complex open-ended responses, the matching process involves the development of a scoring guide for each item, including explanations of what a typical response at each waypoint must contain, as well as well-chosen exemplars, and a training scheme for judges (see Wilson, 2005, or Wilson, in press, for illustrations of these). For example, the following response to Question 1(b)

> When we used the ruler, there were more mistakes (more gaps and laps) when we switched to the meter stick, there were fewer mistakes. So the measurements with the ruler are more spread out than the measurements with the meter stick.

would be mapped to MoV2.

Interlude: reality check. The theory and practice of MoV competence is not sufficiently advanced that a one-item instrument such as Item 1(a) used above could dependably provide information of sufficiently high quality for most intended purposes. Hence, good instrument design demands a strategy where a property such as the MoV competence is sampled across important facets, and this can only be accomplished by using multiple items (Wilson, in press). A more realistic situation is then that instrument developers would create a set of K items $\boldsymbol{I}_m = \{I_{mk}\}$, where $k = 1, ..., K$, each of them designed to interact with a facet of the MoV competence of a student. Then, once the student has responded to the items in the test, the transduced property is the vector $\boldsymbol{X}_m = (X_{m1}, ..., X_{mK})$ of the responses to the set of items. These items may be designed so that some responses are indicative of more sophisticated understanding of variation, i.e., higher Θ, and other responses are indicative of less. The local values for the MoV competence of reader a can then be gathered into a vector $\boldsymbol{x}_i = (x_{i1}, ..., x_{iK})$ of 1s and 0s of length K, which might be summarized by taking the sum of the K values,[25] as before, referred to as "sum-scores", or "total scores", as indeed has been done in psychology and education for over a century within the framework of classical test theory (e.g., Nunally & Bernstein, 1994: p. 215–247, p. 308–310).

Public scale construction and application, and calibration. The local values obtained by the test are not broadly useful other than when using that specific set of items $\boldsymbol{I}_m$, as the relationship between the local values for this set and the local values for a different set of items is not generally known. Whenever the comparability of results may remain local to the context where the same instrument is used (e.g., within a single teacher's classroom), this may not be a problem, but a further step would be needed to equip these local values for public usage, through the definition of a public scale of MoV against which different MoV tests could be calibrated. As Fig. 7.8 shows, in principle this requires identification of an appropriate set of public reference properties, and therefore of publicly available MoV waypoints.

Thus, for each student represented by their local values in the data set, an estimate of the location of her MoV competence on the public scale is obtained, corresponding to a Basic Evaluation Equation, at least for this initial stage of the instrument development. The interpretation of this estimate is aided by using a chart sometimes referred to as a "Wright map". This capitalizes on the most important feature of the output from an analysis using the Rasch model: the estimated locations of the respondents on the property described by the construct map are on the same scale as the estimated locations of the categories of item responses. This allows one to relate the empirical findings about the items to the educational hypotheses embodied in the construct map. Such a feature is crucial for both the measurement theory and measurement practice in a given context:

(a) in terms of theory, it provides a way to empirically examine the waypoints, and adds this as a powerful element in studying the validity of use of an instrument;

[25] Consistently with the hypothesis that some responses are indicative of more sophisticated understanding, it is also possible to use weighted sum here, but the statistical model employed in the analysis reported below does indeed use an unweighted sum for dichotomous items.

(b) in terms of practice, it allows the measurers "go beyond the numbers" in reporting measurement results to practitioners and consumers, and equip them to use the construct map as an important interpretative device.

As reported in Wilson and Lehrer (2021), results from the analysis of data were collected using a sample of about one thousand middle school students from multiple school districts to calibrate the MoV items, by fitting a partial credit calibration model (a one-dimensional Rasch-family item response model (Masters, 1982)) to the item responses. The resulting Wright map for this analysis is shown in Fig. 7.20: looking from left to right, the first column shows the logit scale, the second shows the distribution of the students as a histogram (in horizontal orientation), running from lowest ability at the bottom to highest at the top, and the raw sum-score approximately corresponding to the histogram bars is shown next; the next five columns show the thresholds[26] for each of the MoV items, separated out into those that relate to each waypoint, with the waypoints indicated along the bottom of each column.[27]

[26] The *k*th threshold is the item parameter located at the point where the probability of a response in category *k* or below equals the probability of a response above *k*, and hence both equal 0.50.

[27] Note that a lower waypoint was added after initial data collection, having no link to the original construct map and representing students who only did use relevant phrases and words, hence referred to as "NL(ii)", the "ii" indicating that there was an even lower level of coding, referred to as "NL(i)", indicating students who actually did not respond.

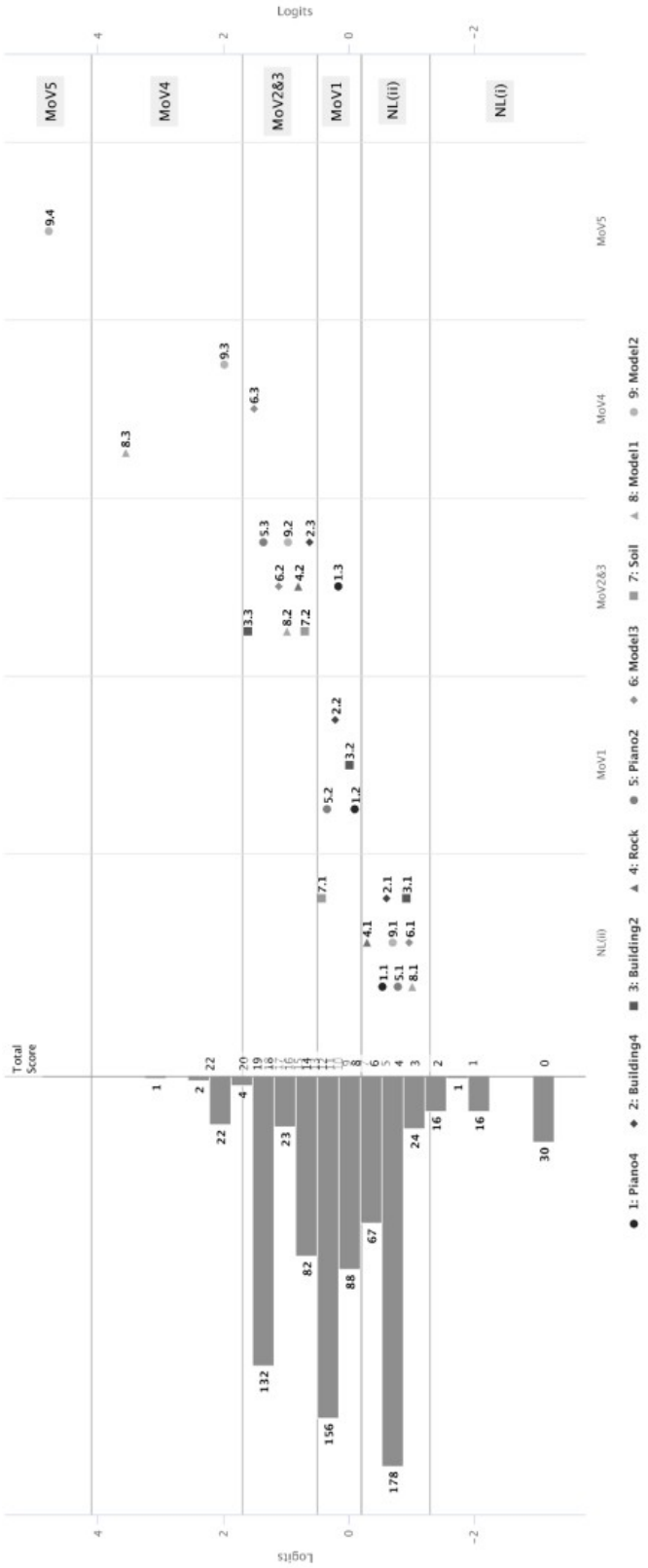


Fig. 7.20 A Wright map for MoV competence (the location of the threshold for each item is represented *i.k*, where *i* is the item number and *k* is the item score; for example, 9.2 is the second threshold location for item 9, i.e., the threshold between the scores 0 and 1 compared to scores 2 to 4, the maximum for item 4)

The last column on the right-hand side shows the *banding* of the Wright map. Note that the waypoint for each item is represented by a single location on the logit scale and hence the set of thresholds for each waypoint has multiple locations: this raises the issue of how best to represent the waypoint. The solution is to use the range of that set of locations as representing each waypoint, and we call that range the *band* for that waypoint. This marks the transition from an ordered qualitative set of waypoints on the construct map to an empirical "band" on the logit scale. Thus, in Fig. 7.20, parts of the logit scale (the bands) are delineated that correspond to the set of thresholds for each waypoint. Due to the uncertainties and many influencing properties inherent in the item design process, it is not always possible to do so in a completely non-overlapping way, and it can be seen here that, although the bands are mainly inclusive of the relevant thresholds, some are not contained in their respective bands (e.g., thresholds 7.1 and 1.3).

Thus, the MoV competence $\Theta[a]$ under measurement can now be associated with public values θ_j via the Basic Evaluation Equation using the scale based on, and interpreted through, the public references $\Theta[s_j^*]$, corresponding to the levels in the construct map in Fig. 7.17.

Conceptualizing the BEAR Assessment System using the Hexagon Framework. In the remainder of this section we discuss how the perspective of the Hexagon Framework can be exploited to inform an instrument development process used in psychosocial measurements. Our aim is to show how the Hexagon Framework can be seen as consistent with, and hence underlying the logic of, a specific instrument design approach that is used in the psychosocial sciences, the *BEAR Assessment System* (BAS: Wilson, 2005; in press). Thus, the BAS is described here as an overlay on the Hexagon Framework: we see this as an illustration of how the Framework can be used not only as a philosophical foundation for measurement, but also as a basis for other useful scientific conceptualizations and models.

The *Data Modeling* project used the BAS to develop an instrument to measure MoV competence. The BAS uses four "building blocks" to develop an instrument: (a) the construct map, (b) the items design, (c) the outcome space, and (d) the Wright map. These building blocks are used in an iterative sequence during instrument development. For a more detailed account of an instrument development process that works through these building blocks, see Wilson (2005; in press). Below, each of these building blocks is considered in turn with respect to MoV competence and related to the Hexagon Framework.

The *initial* construct map for the MoV competence is similar to the final one, shown in Fig. 7.17, but for one important difference, which is elaborated on below.

During the *Data Modeling* project, the typical work-pattern was to (a) develop instructional materials and practices based on the current idea about MoV (e.g., starting with the initial ideas about the construct map based on the literature), (b) try them out in classrooms, and with input from the teachers involved, and (c) revise them, including updating the construct map. This was repeated over several different iterations in different classrooms with different teachers. At the initial stage, the waypoints were ordinally related to one another. The next few steps lay out how the project team worked to find a quantitative relationship between them. The waypoints in the construct map are merely labeled here: the interested reader can find detailed descriptions in Wilson and Lehrer (2021).

As should be clear from the discussion above, the place of the construct map in the Framework is as a description of the property under measurement (see the construct map ellipse in Fig. 7.21), and as such it is analogous, in the example of the development of thermometers, to a sequence of fixed points of temperature, like the freezing point and the boiling point of water.

The next step in instrument development under the BAS is the items design, which involves developing ways in which the property described by the construct map could be manifested via an empirical situation: this is the transduction in the Hexagon Framework (see the items design ellipse in Fig. 7.21). For example, the *Data Modeling* items often began as everyday classroom experiences and events that teachers have found to have a special significance in learning of variability concepts. Example items are shown in Fig. 7.18, in the *Piano Width* task.

The next step under the BAS is the development of the outcome space, which involves the step from the classification of the population of potential student responses to the assignment of scores for use in the statistical analysis to follow. To facilitate this, the project collected a sample of student responses during the instrument development to support the instrument development. Thus, this is the equivalent in the Hexagon Framework of the steps from the transduced properties to the local values (see the outcome space ellipse in Fig. 7.21).

The next step in this developmental process is to relate these scores back to the construct map. This is initiated through the fourth BAS building block, the Wright map (see the Wright map ellipse in Fig. 7.21). As mentioned above, a statistical calibration model is used to analyze the resulting data.

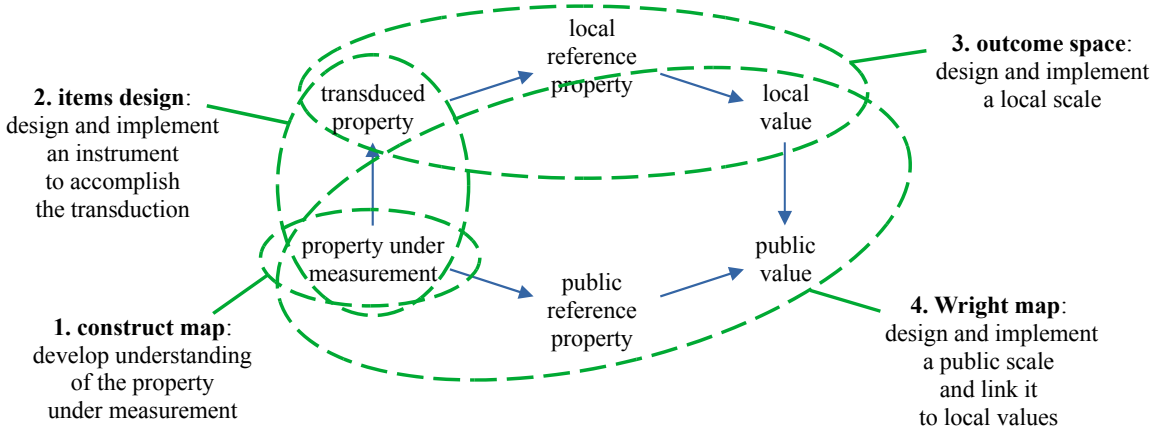


Fig. 7.21 The Bear Assessment System (BAS) building blocks overlaid on the Hexagon Framework

The most striking feature of the banded Wright map in Fig. 7.20 is that the waypoints for MoV2 and MoV3 were found to occupy the same segment of the logit scale. In the initial investigations of this Wright map, it was found that the thresholds for the waypoints Mov2 and Mov3 were thoroughly mixed. A large amount of time was spent exploring this, both quantitatively, using the data, and qualitatively, examining item contents, and talking to curriculum developers and teachers about the apparent anomaly. The conclusion was that for these two waypoints, although there is certainly a necessary hierarchy to their lower ends (i.e., there is little hope for a student to successfully use a chance-based device such as a spinner to represent a source of variability (MoV3) if they cannot informally describe such a source (MoV2)), these two waypoints can and do overlap quite a bit in the classroom context. Students are still improving on MoV2 when they are initially starting on Mov3, and they continue to improve on both at about the same time. Hence, at least formally, while it was decided to uphold the distinction between Mov2 and MoV3 in terms of content, it also seemed best to ignore the difference in difficulty of these waypoints, and to label the segment of the scale (i.e., the relevant band) as “MoV2” and “MoV3” as in the final construct map shown above in Fig. 7.17. This is an example of how the empirical findings during the measurement development process can lead to modifications of the understanding of the property under measurement, and it is also an illustration of

the iterations of the BAS: in effect, the researchers involved in the project changed the definition of the property under measurement as they worked around the Framework!

Thus, the building blocks of the BAS can be seen as an overlay on the Hexagon Framework and hence incorporating the logic of the Hexagon Framework. The logit scale is a representation of the property under measurement in this situation, and the banded Wright map is the feature that affords the possibility of a criterion-referenced interpretation of the students estimated locations on the logit scale in terms of the MoV waypoints, which serve as the public reference properties for MoV competence.

The account above has traced the progress of the instrument development through just one iteration of the BAS, but in fact it is just one embedded in several. There were several initial partial circuits around the BAS where the measurement developers were working closely with the curriculum developers and the co-operating teachers to develop not only the items and the outcome space, but also even the waypoints themselves. In addition, a final run-through of the data collection and a second analysis were needed to provide more items for the instrument and a better, more representative sample of the student population.

The idealized model of direct measurement described in the preceding sections needs to be generalized to the more realistic case in which uncertainties are taken into account.

7.4 Measurement quality according to the model

As operationalized by the Hexagon Framework, the Basic Evaluation Equation needs to be augmented with an assessment of the quality of the information conveyed by measurement. As discussed in Sect. 3.2, this role is played by measurement uncertainty, which is inversely related to information quality: the better the quality the less the uncertainty. By taking uncertainty into account, the model of direct measurement introduced above is improved and generalized. In the human sciences, the quality of measurement is usually assessed in terms of validity (see Sect. 4.3) and reliability (see Sect. 3.2.1): while reliability is usually assessed via a quantifier, this is seldom the case for validity, and if so, then it would only be quantified for some components of what is generally termed “validity”.

The acknowledgment of the structural, and not only operational, importance of measurement uncertainty is relatively new in physical metrology: “the need to find an agreed way of expressing measurement uncertainty in metrology” was stated in the Recommendations issued by the International Committee of Weights and Measures (CIPM) in 1980-81 (quoted in JCGM, 2008). In the human sciences, the need to investigate the validity of the measurement has been a basic element of measurement practice since the early 20th century (see Sect. 4.3). This can be interpreted as a revision of the basic black box model: given an input property, a measurement is expected to produce not only one or more values but also some information on the quality of the information that such values provide on the measured property.

As mentioned, until a recent past, measured values were reported together with estimates of measurement *errors*. The relation between error and uncertainty in measurement is complex: uncertainty has sources that are not what would traditionally be described as errors, as in the case of definitional uncertainty, and some errors could be known only with some uncertainty; hence error not only may generate uncertainty, but also may have its own uncertainty.[28] More importantly, the

[28] Complicating matters, some authors refer to error and uncertainty interchangeably (Kirkup & Frenkel, 2006), as noted by Taylor: “In science, the word error does not carry the usual connotations of the terms mistake or blunder. Error in a scientific measurement means the inevitable uncertainty that attends all measurements. [... Here] error is used

emphasis on uncertainty is a result of a conceptual shift in the recent metrological literature from a purely empirical to a model-based approach, incorporating both empirical and informational interpretations of measurement. As a consequence, the central concept of measurement science is arguably no longer the "true value" that exists independently of measurement and that would be obtained by an error-free empirical process. While measurement is still sometimes characterized as a process aimed at estimating the true value of a property (a prominent example is in Possolo, 2015: p. 12), the very idea of a measured property having an inherent true value requires clarification as soon as the unavoidable role of models in measurement is accepted. For example: does the true value of a property change if the model of the property or the model of the measurement change? is therefore a true-in-a-model value? Does the true value remain unique also in presence of non-null definitional uncertainty? Or are there in this case multiple true values? See the related discussion in Box 6.1.

Hence, in what follows we do not deny in principle the hypothesis that properties have a true value, but neither do we rely on it: instead, we attempt to provide an encompassing standpoint which should be understandable and acceptable independently of this hypothesis. Like the rest of this chapter, and like most of this book, what follows may be read as starting from the VIM definition of <measurement> as a process of *reasonable* attribution of values to properties of objects (JCGM, 2012: 2.1) and aimed at establishing sufficiently well-defined criteria of such reasonableness. An appropriate characterization of measurement uncertainty plays a key role in service of this goal. What follows may be interpreted as a reconsideration of the basic components of measurement uncertainty, as introduced in Sect. 3.2.4, in light of the model presented above. But, first of all, we need to reconsider the Hexagon Framework and expand it in order to take into account the possibility of feedback with a measuring instrument.

7.4.1 Measurement that involves feedback

We have assumed so far that the interaction between the object under measurement and the measuring instrument, as realized in the transduction stage (see Sect. 7.3.2), is unidirectional: the interaction changes the state of the instrument, and therefore the transduced property X_m, as modeled by the map $\Theta[a] \rightarrow X_m$. In a more general case, however, the interaction produces a change also in the state of the object under measurement. In particular, when the object under measurement is a human being (or a set thereof), as is usually the case in the human sciences, the objects under measurement may be aware of their being objects under measurement. In this circumstance, not only might interaction uncertainty become a critical component of the uncertainty budget, but the structure of the process itself becomes more complex due to the presence of one or more feedback loops.

Three structural cases may be identified, as follows.

First case (*no feedback*): these are measurements in which the interaction with the measuring instrument does not induce a change in the measurand (an obvious example is the measurement of the spectral density of the radiation emitted by a star: of course, the state of the star is not affected by the operation of the spectrometer). In this case there is no feedback in the process, and therefore there are no problems in objectivity resulting from the one-way interaction.

Second case (*non-oriented, random feedback*): these are measurements in which the interaction with the measuring instrument induces a random change in the measurand, for example due to an uncontrolled transfer of energy between the instrument and the object under measurement. In this case

exclusively in the sense of uncertainty, and the two words are used interchangeably." (1997: p. 3). More or less explicitly, this denies that there is anything new in what has happened on this matter in the last decades, as witnessed in particular by the publication in 1993 of the *Guide to the expression of uncertainty in measurement* (GUM) (JCGM, 2008).

a non-oriented feedback is present in the process: in usual conditions, the measurement trueness is not affected by this loop (in other words, no systematic errors arise from the interaction), and some problems in objectivity may arise due to insufficient measurement precision, revealed by large random errors.

Third case (*oriented, non-random feedback*): these are measurements in which the interaction with the measuring instrument induces a non-random change in the measurand (which is called a "loading effect" in the context of electrotechnology, for example). Whenever such a change is identified and modeled, typically as a systematic error (or as a bias, i.e., the estimate of systematic error; JCGM, 2012: 2.18), its effects may be experimentally minimized or mathematically corrected. As in the previous case, this situation of oriented feedback generally results in problems in objectivity. In the human sciences, this is the context in which the so-called "Hawthorne effect" (Landsberger, 1958) arises, in which individuals alter their behavior due to their being aware of being observed. An undesired consequence of this effect is summarized in what is sometimes called Goodhart's law: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart, 1981).[29] A related phenomenon in educational measurement is sometimes called "teaching to the test", when the awareness of educators and their students of being tested (and the consequences thereof) alters the nature of teaching and learning. A peculiar consequence is that, at least in the short term, measurement-like activities may be exploited as managerial tools, leading individuals to change their behaviors only because they are informed that some measurements will be performed on them, even if the measurements are ultimately not utilized or even performed, in which case the feedback loop does not actually include the measuring instrument.

7.4.2 Uncertainties in the stages of direct measurement

Our updated account of measurement uncertainty begins with the consideration of the preliminary stages of any actual measurement: the definition of the measurand and the calibration of the instrument.

Regarding the definition of the measurand. Measurement is aimed at acquiring the information required to assert a Basic Evaluation Equation $\Theta[a] = \theta$, i.e., the equality of the property Θ of the object a and the public value θ (see the analysis in particular in Sects. 5.1.1 and 6.4). In the simplest condition, the measurand is defined as the property which interacts with the instrument and triggers the transduction, so that, in a (tautological) sense, we measure what the instrument measures (this corresponds to what in Box 2.3 is called an operational strategy of measurand definition). This might be acceptable when the information acquired by measurement is required only in the specific time and place in which the measurement is performed, and thus the model of the measurand can remain implicit: the intended property is the same as the effective property, and nothing else needs to be specified. In the example of the measurement of the temperature Θ of an object a, defining the measurand simply as the temperature that is transduced by the thermometer implies that one can assume that neither the thermal inhomogeneities of the object (different points of the object might have different temperatures) nor the environmental conditions (e.g., the temperature of the object

[29] Several cases of this phenomenon are proposed by Jerzy Muller in his book so explicitly titled *The tyranny of metrics*. An example: "In England, in an attempt to reduce wait times in emergency wards, the Department of Health adopted a policy that penalized hospitals with wait times longer than four hours. The program succeeded – at least on the surface. In fact, some hospitals responded by keeping incoming patients in queues of ambulances, beyond the doors of the hospital, until the staff was confident that the patient could be seen within the allotted four hours of being admitted." (2018: p. 5).

might be affected by the pressure, or humidity of the environment) are relevant in the information to be reported.

However, in scientific measurement, a Basic Evaluation Equation is generally expected to convey widely transferable information. Values, on the right-hand side of the equation, can in principle be interpreted in the same way everywhere and always thanks to their metrological traceability (see Sect. 3.3.1). Analogously, measurands, on the left-hand side of the equation, should be interpretable beyond the here-and-now situation. This is accomplished by explicitly defining the measurand, by means of a model which identifies the measurand by description instead of by purely indexical means, and therefore by taking into account the possible differences between the measurand and the property that produces the transduction: the information is empirically acquired about the effective property but is reported about the intended property, and a purpose of the model is to establish a connection between these two. Sometimes these differences can be considered explicitly, if the model has a mathematical form in which a value of the intended property is calculated as a function of both the effective property and the appropriate corrections, as when the measurand is the temperature of an object in given environmental conditions but the temperature is measured in different conditions, and there is a known law connecting such environmental conditions to the property under measurement.

But, as usual, there is a price to be paid for improving the transferability of the measurement information: the greater the specificity of the information, the greater its uncertainty, which in this case is uncertainty about the definition of the measurand, what the VIM calls the *definitional* uncertainty (JCGM, 2012: 2.27) (see Sect. 3.2.4). Thus, ignoring the distinction between the intended and the effective property amounts to the elimination of definitional uncertainty from the model. For example, considering the temperature of water in a container, the effective property is the temperature of that part of the water with which the thermometer interacts, in the context of the, possibly unknown, conditions of the water in the container at the time of the interaction. However, the measurand could be defined by a specification of the conditions of the object (i.e., the water) and the environment (i.e., the container and the surrounding space), for example by assuming that the water is thermally uniform and the measurement takes place at a given environmental pressure. Assuming this makes the information more transferable, but at the cost of non-null definitional uncertainty: we must take into account the differences between the *specified* conditions (i.e., "the water is thermally uniform and the measurement takes place at a given environmental pressure") and the *actual* conditions of the interaction of the object and the measuring instrument.[30] The place of definitional uncertainty in the Framework is depicted in Fig. 7.22. As noted in Sect. 3.2.4, there are many types of definitional uncertainty in the context of measurement in the human sciences (sometimes referred to as "threats to validity"). One particular threat is *construct underrepresentation*: this is where the effective property is less complex or rich than the intended property. An example in the case of RCA would be, for example, where the property is considered to pertain to comprehension across paragraphs or texts, but the transducers (i.e., the reading comprehension items) are instead always focused on specific words or phrases within a sentence.

[30] The measurand definition and its influence on the design and the operation of measurement are subjects that still require investigation – see, e.g., Baratto (2008) and Morawski (2013).

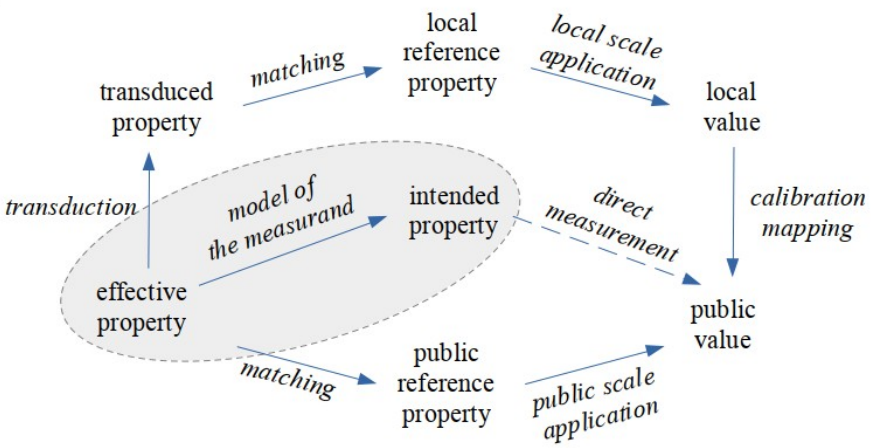


Fig. 7.22 An extension of the Hexagon Framework including the distinction between intended property and effective property, and the place (highlighted by the gray ellipse) of definitional uncertainty in it; the mapping labeled "direct measurement" is the operationalization of a Basic Evaluation Equation

Regarding the definition and dissemination of the public scale and calibration. We assume that at a given time the instrument has been calibrated against some reference objects, i.e., measurement standards, with reference properties $\Theta[s_j^*]$ and corresponding values θ_j. This requires, first, that the public scale $\Theta[s_j^*] \rightarrow \theta_j$ was effectively disseminated, from the primary realization of the definition of the reference properties (and therefore of the unit, in the case of quantities), via a metrological traceability chain. Thus, along the chain and across contexts, inaccuracies and instabilities may affect the reproduction of the reference properties in such a way that the properties of the primary standards may differ from the properties $\Theta[s_j^*]$ of the working standards. This leads to uncertainties in the public scale $\Theta[s_j^*] \rightarrow \theta_j$ used for the instrument calibration. Moreover, calibration requires some empirical processes to be performed, i.e., transduction and matching, in which influence properties may affect the instrument indication and therefore the construction of the local scale $X_i^* \rightarrow x_i$. These uncertainties in both the public scale and the local scale combine to form a *calibration uncertainty* affecting the calibration map $x_i \leftrightarrow \theta_j$, as depicted in Fig. 7.23. As noted above, the creation of public reference objects in the human sciences may be based on the means of sum-scores of specified groups (e.g., Grade 6 students in a given school system taking an RCA test). This can lead to calibration uncertainty if the means change over time, and these changes are not accounted for in the public scale, as with the Flynn effect (Flynn, 1987).

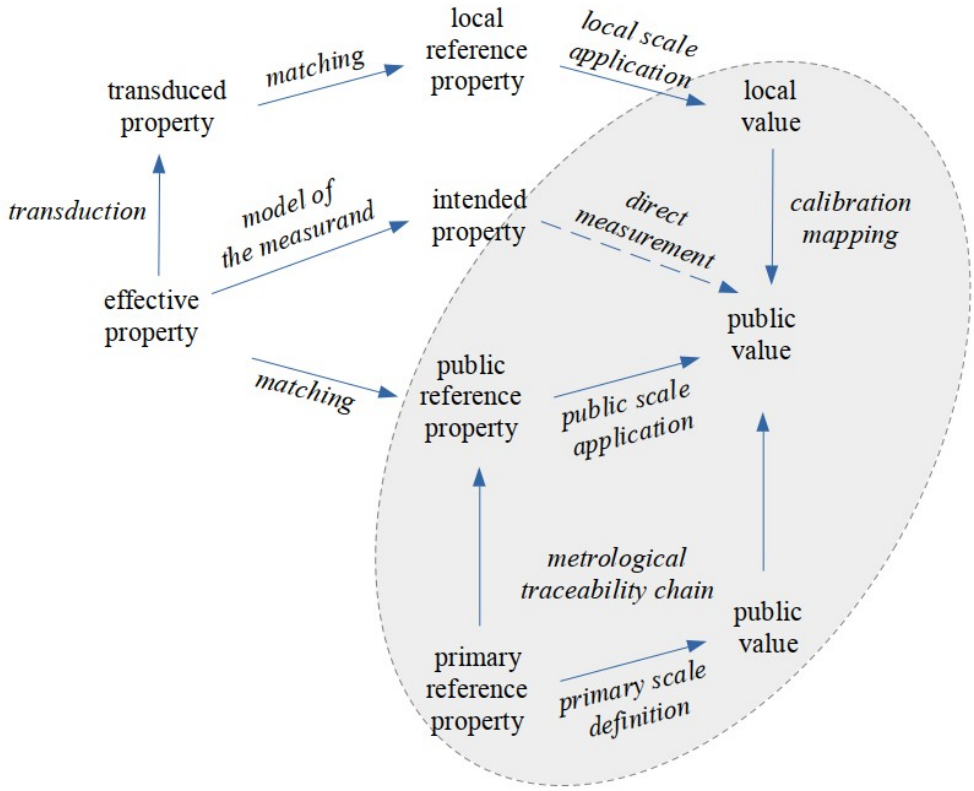


Fig. 7.23 The Hexagon Framework including the definition of the primary scale and its dissemination, and the place (highlighted by the gray ellipse) of calibration uncertainty in it

Regarding transduction and matching. As with any empirical process, the specifications in the measurement procedure will not be entirely realized when the measuring instrument is operated. This is due, first, to the unavoidably limited stability and selectivity of the transducer. Moreover, there could be errors in matching the transduced property X_m with respect to the reference properties X_i^* in the local scale (e.g., the so-called "reading errors" in the case of instruments with analog scales). In addition, the local scale could also be affected by instabilities, resulting in a time-dependent mapping $X_i^* \rightarrow x_i$, which is only uncertainly known. These issues lead to a non-null *instrumental uncertainty*. One example for RCA would be where the RCA test is simply mis-scored, whether by human or machine.

Finally, the fact that the object under measurement must somehow interact with the transducer may induce an unwanted change in the state of the object, which in turn produces an *interaction uncertainty*, as depicted in Fig. 7.24. An example of this in RCA testing would be where the text passage that was used contained elements that affected (in either positive or negative ways) some individual readers (in either positive or negative ways), which altered their responses to items.

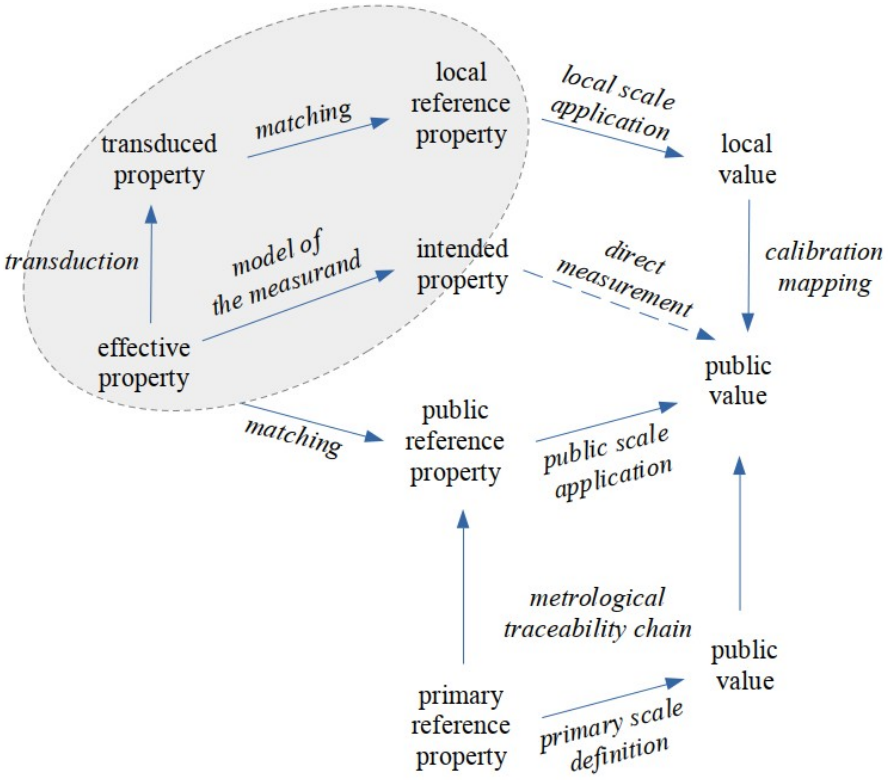


Fig. 7.24 The place (highlighted by the gray ellipse) of instrumental uncertainty and interaction uncertainty in the Hexagon Framework

In summary, the different components of the measurement model invoke different aspects of measurement uncertainty, as depicted in Fig. 7.25. As discussed in Sect. 3.2.5, where these can be quantified, these components can be gathered into an uncertainty budget.

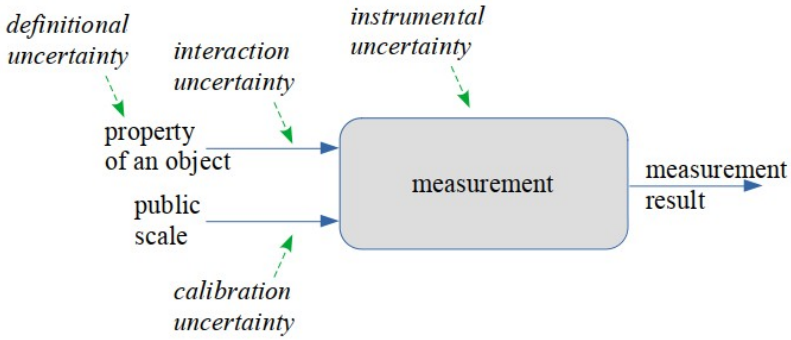


Fig. 7.25 An updated black box model, now including the components of measurement uncertainty (an updated version of Fig. 3.3)

7.4.3 Quality of measurement as objectivity and intersubjectivity

The overall understanding of the quality of measurement and its results may be characterized in terms of the two basic features, which we have called *object-relatedness* and *subject-independence* (Mari et al., 2012; Maul et al., 2018). We introduce them here and show how, in the context of the Hexagon Framework, they are related to the different kinds of measurement uncertainty discussed in the previous section.

Object-relatedness ("objectivity" for short) is the extent to which the conveyed information is about the measurand, i.e., the intended property, and nothing else. According to the model we have presented, the problem of objectivity of measurement and its results is threefold.

First, considering definitional uncertainty, empirical properties are interrelated because they are mutually dependent, so that the measurand depends on other properties, i.e., the affecting properties, as discussed in Sect. 7.2.2. Since the information produced by measurement is supposed to be transferable, i.e., usable not only in the time and place in which it was obtained, and not all the relevant properties might be known at that moment, the issue arises of defining the measurand in a sufficiently specific way for making the information transferable without losing the reference to the measurand. *Definitional uncertainty* is the related component of uncertainty, and where it is quantifiable, it can be incorporated into the uncertainty budget.

Second, considering interaction uncertainty, while the object under measurement needs to change the state of the measuring instrument – thus resulting in a transduced property – for a measurement to take place, the opposite effect also sometimes happens, with the consequence that the object under measurement also changes its state, and then possibly the property under measurement changes in turn, due to its interaction with the instrument, as discussed in Sect. 7.4.1. The result is a loss of objectivity, which may be quantified by *interaction uncertainty*.

Third, considering instrumental uncertainty, the measuring instrument is generally sensitive not only to the measurand but also to other properties (the influence properties, as discussed in Sect. 7.2.2), with the consequence that its output depends also on such properties: since the information produced by measurement is supposed to be usable independently of the instrument by which it was obtained, the issue arises of characterizing the instrument behavior in a sufficiently specific way for making it possible to extract information on the measurand by filtering out the spurious information generated by influence properties. In Sect. 3.2.1 the metrological behavior of an instrument is characterized in terms of its accuracy, and then more specific features such as trueness and precision. When reported in terms of measurement results, this component of objectivity may be quantified by means of *instrumental uncertainty*.

Subject-independence ("intersubjectivity" for short) takes into account the goal that the conveyed information be interpretable in the same way by different persons in different places and times. This requires that the information produced by measurement is reported in a way that is independent of the specific context and only refers to universally accessible entities, so that in principle its meaning can be unambiguously reconstructed by anyone. Metrological systems, including quantity units realized by measurement standards disseminated through traceability chains, are developed and maintained to fulfill this requirement. The appropriate calibration of the measuring instrument guarantees the metrological traceability of the information it produces, and therefore the condition of intersubjectivity. *Calibration uncertainty*, which includes all uncertainties related to the definition of the public scales and their realizations in the measurement standards in the traceability chain, is then what may quantify intersubjectivity.

The characterization of measurement in terms of objectivity and intersubjectivity[31] is relevant for both users and designers:

[31] Sometimes the distinction between objectivity and intersubjectivity is not maintained, and they are conflated in a single concept of <non-subjective>. An explicit example is: "A highly disciplined discourse helps to produce knowledge independent of the particular people who make it. This last phrase points to my working definition of objectivity. It is, from the philosophical standpoint, a weak definition. It implies nothing about truth to nature. It has more to do with the exclusion of judgment, the struggle against subjectivity." (Porter, 1995: p. ix).

- users are generally interested in the information, not the way it is produced; from the user's point of view, objectivity and intersubjectivity are features of the products of the process, i.e., of measurement results;
- designers are interested in the way the information may be produced; from the designer's point of view, objectivity and intersubjectivity are features of the process, then inherited by its products, i.e., first of all of measurement.

Hence, we can see objectivity and intersubjectivity as features of both the process (i.e., measurement) and the products (i.e., measurement results).[32]

As they have been characterized, objectivity and intersubjectivity are embedded in measuring systems: in other words, measuring systems are designed, set up (including via their calibration), and operated so to be able to produce information with the expected degree of objectivity and intersubjectivity, i.e., able to produce measurement results with a measurement uncertainty which is less than *target* uncertainty, the "upper limit [of uncertainty] decided on the basis of the intended use of measurement results" (JCGM, 2012: 2.34). This highlights the pragmatic nature of measurement: what counts as high or low quality is relative to the purpose of the measurement; if a comparatively lower quality instrument provides results of sufficient accuracy, using it could lead to still acceptable (and cheaper) measurements.

7.4.4 Can measurement be "bad"?

According to the characterization we have just proposed, objectivity and intersubjectivity are independent features: something can be objective but not intersubjective (as might happen in the case of the usage of an uncalibrated measuring system), or vice versa (as when the result of an evaluation is expressed in the customary format for the values of quantities, i.e., number times unit, but was obtained through a badly flawed measurement). Together they identify the two dimensions of quality of measurement: the claim of the possibilities *of obtaining information about empirical properties*, and *of socially reporting such information*.[33] It is thus through their objectivity and intersubjectivity that measurement results are considered to be of good quality. Since objectivity and intersubjectivity are not Boolean (i.e., yes-no) conditions, in a given operational situation one could set a threshold of minimum acceptable objectivity and intersubjectivity, aimed at guaranteeing that the results of measurement will be useful for their intended use. This highlights the pragmatic nature of measurement: the same measurement results might be considered good for some purposes and bad for some others. Hence, objectivity and intersubjectivity are features of good measurements, not of measurement as such. This allows <bad measurement> to be an acceptable concept – i.e., not all measurements are good –, where "bad" is meant as <not sufficiently objective and intersubjective according to the given purposes of the measurement>.[34]

[32] As noted above, the concept of validity in human sciences measurement is closely related to definitional uncertainty. However, it is an expansive concept, and as such it has aspects that also overlap with other parts of objectivity and intersubjectivity.

[33] Interestingly, these dimensions correspond to the two main stages of a measurement process of (a) transduction and matching and (b) calibration mapping, as connected by the local scale application. The idea that measurement is to be modeled on the basis of such two stages is not unusual, though the terms may be different. For example, Roman Morawski calls them "conversion" and "reconstruction" (2013) and Giovanni Battista Rossi and Francesco Crenna call them "observation" and "restitution" (2018).

[34] This implies that <measurement> cannot be defined in terms of objectivity and intersubjectivity, as instead sometimes suggested (e.g., "the result of measurement must meet the condition of objective truth", Piotrowski, 1992: p. 1), also, mistakenly, by one of the present authors (Mari et al., 2012: p. 2109).

The objectivity and intersubjectivity of measurement results may be interpreted as their overall "degree of quality", which is (inversely) specified and quantified by measurement uncertainty: a good measurement produces measurement results whose uncertainty is

- beyond the definitional uncertainty of the measurand (again, a measurement uncertainty less than definitional uncertainty corresponds to a waste of resources devoted to design and perform the measurement), but
- less than the specified target uncertainty (a measurement uncertainty greater than target uncertainty corresponds to a useless measurement).

An emphasis on sufficient objectivity and intersubjectivity for a given purpose is then operationally useful, for the general guidelines it provides regarding the design and performance of measurements (e.g., in Petri et al., 2015), but it is still too specific at least in one respect: it would assume that measurement is always good measurement. While pragmatically this is sound – if we know that what we are doing is a bad measurement we (hopefully) avoid doing it – the concept 'bad measurement' as such is not contradictory, and bad measurements do not fulfill the condition of sufficient objectivity and intersubjectivity. In other words, in order to maintain the VIM's characterization of "reasonableness", objectivity and intersubjectivity are useful but still not sufficient: some other conditions have to be identified. This, among other things, is discussed in the next chapter.

References

Baratto, A. C. (2008). Measurand: a cornerstone concept in metrology. *Metrologia, 45*, 299–307.

Bich, W. (2008). How to revise the GUM?. *Accreditation Quality and Assurance, 13*, 271–255.

Boumans, M. (2007). Invariance and calibration. In M. Boumans (ed), *Measurement in economics: A handbook* (pp. 231–248). London: Academic Press.

Campbell, N. R. (1920). *Physics – The elements*. Cambridge: Cambridge University Press.

Cartwright, N. (1999). *The dappled world – A study of the boundaries of science*. Cambridge: Cambridge University Press.

Chang, H. (2007). *Inventing temperature – Measurement and scientific progress*. Oxford: Oxford University Press.

Cook, A. H. (1994). *The observational foundations of physics*. Cambridge: Cambridge University Press.

Dray, A. J., Brown, N. J. S., Diakow, R., Lee, Y., & Wilson, M. (2019). A construct modeling approach to the assessment of reading comprehension for adolescent readers. *Reading Psychology, 40(2)*, 191–241.

Ellis, B. (1968). *Basic concepts of measurement*. Cambridge: Cambridge University Press.

Flynn J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*, 123–149.

Gertsbakh, I. (2003). *Measurement theory for engineers*. Berlin: Springer.

Giordani, A., & Mari, L. (2019). A structural model of direct measurement. *Measurement, 145*, 535–550.

Goodhart, C. (1981). Problems of Monetary Management: The U.K. Experience. In A. S. Courakis (Ed.), *Inflation, Depression, and Economic Policy in the West* (pp. 111–146). Totowa, NJ: Barnes and Noble.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge: Cambridge University Press.

International Electrotechnical Commission (IEC) (various publication dates). *International Electrotechnical Vocabulary (IEV)*. Geneva: IEC. Online: www.electropedia.org

International Standardization Organization (ISO) and other three International Organizations (1984). *International Vocabulary of Basic and General Terms in Metrology (VIM)* (1st ed.). Geneva: International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Organization for Standardization (ISO), International Organization of Legal Metrology (OIML).

Joint Committee for Guides in Metrology (JCGM) (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (JCGM) (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.). Sèvres: JCGM (2008 version with minor corrections). Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2012). *JCGM 106:2012, Evaluation of measurement data – The role of measurement uncertainty in conformity assessment*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Kirkup, L., & Frenkel, R. B. (2006). *An introduction to uncertainty in measurement using the GUM (Guide to the expression of Uncertainty in Measurement)*. Cambridge: Cambridge University Press.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Kyburg, H. E. (1984). *Theory and measurement*. Cambridge University Press.

Landsberger, H. A. (1958). *Hawthorne Revisited*. Ithaca, NY: Cornell University Press.

Lira, I. (2002). *Evaluating the measurement uncertainty – Fundamentals and practical guidance*. Bristol: IOP Publishing.

Mari, L. (2005). The problem of foundations of measurement. *Measurement, 38*, 259–266.

Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: a pragmatic view. *IEEE Transactions on Instrumentation and Measurement, 61*, 2107–2115.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–121.

Masters, G., & Forster, M. (1997). *Mapping literacy achievement: Results of the 1996 National School English Literacy Survey*. Hawthorn, Australia: ACER Press.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement, 116*, 611–620.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments* (CSE technical report). Washington, DC, and University of California at Los Angeles: Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing.

Morawski, R. Z. (2013). An application-oriented mathematical meta-model of measurement. *Measurement, 46*, 3753–3765.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton: Princeton University Press.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction*. New York: Harcourt Brace Jovanovich.

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Palmer, de Forest A. (1912). *The theory of measurements*. New York: McGraw-Hill.

Petri, D., Mari, L., & Carbone, P. (2015). A structured methodology for measurement development. *IEEE Transactions on Instrumentation and Measurement, 64*, 2367–2379.

Piotrowski J. (1992). *Theory of physical and technical measurement*. Amsterdam: Elsevier.

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

Possolo, A. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST measurement results*. NIST Technical Note 1900. Retrieved from www.nist.gov/publications/simple-guide-evaluating-and-expressing-uncertainty-nist-measurement-results

Roberts, F. S. (1979). *Measurement theory with applications to decision-making, utility and the social sciences*. Reading, MA: Addison-Wesley.

Rosen, R. (1978). *Fundamental of measurement and representation of natural systems*. New York: North-Holland.

Rossi, G. B., & Crenna, F. (2018). A formal theory of the measurement system. *Measurement, 116*, 644–651.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/measurement-science

Taylor, J. R. (1997). *An introduction to error analysis – The study of uncertainties in physical measurements* (2nd ed.). Sausalito: University Science Book.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum (now published by Taylor and Francis, New York).

Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice, 37(1)*, 5–20.

Wilson, M. (in press). *Constructing measures* (2nd Edition). New York: Routledge.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wright, B. D., & Masters, G. N. (1981). *Rating Scale Analysis*. Chicago: MESA Press.

Chapter 8.
Conclusion

This chapter aims to conclude the book by providing a high-level interpretation of measurement and its characterizing features. We first develop a semiotic perspective on the information gained via measurement, specifically, how, in any measurement process, syntactic information (i.e., data, in the form of indication values) grounds semantic information (in the form of measurement results), which in turn grounds pragmatic information (in the form of measurement results together with the contextual information that enables decision making). We then briefly retread the path followed in this book, describing how we began with a minimal set of necessary conditions for measurement and then progressively explored issues critical to the development of complementary sufficient conditions, related in particular to the ontology and epistemology of measured properties, the nature of scales, measurability, and measured values, and the roles of empirical and informational processes in measurement. This culminates in a general model of a measurement process that emphasizes the importance of evaluating the quality of the information produced by measurement in terms of object-relatedness ("objectivity") and subject-independence ("intersubjectivity"). We conclude with an argument that, despite differences in subject matter and application, any measurement process can be characterized as an empirical and informational process that is designed on purpose, whose input is an empirical property of an object, and that produces explicitly justifiable information in the form of values of that property.

8.1 Introduction

Measurement is a designed, intentional (not spontaneous, not natural) process – that as such needs to comply with a procedure – performed for improving the previously available information about a property of an object: indeed, while bad and useless measurements are certainly possible, as discussed in Sect. 7.4.4, the results of a measurement are expected, in principle, to be able to improve the available information. This final chapter is devoted to better exploring the claim that *measurement is designed for producing useful information on the measurand*: this will allow us to identify the sufficient condition to characterize measurement that we have been seeking from the beginning of this book.

The model of direct measurement presented in Chap. 7 is the basis of the analysis that follows, in which measurement is interpreted as a tool for communication: first measurement produces *syntactic information* (corresponding to local values, i.e., values of instrument indication), which is then transformed into *semantic information* (corresponding to measurement results), which – if the measurement was appropriately designed and performed – is finally transformed into *pragmatic information*, and therefore information useful for the purposes that led to the design and performance of the measurement itself. This interpretation embeds measurement in semiotics, to which we first provide a short introduction.

8.1.1 Syntactic, semantic, and pragmatic information

The distinction among syntax, semantics, and pragmatics (see Fig. 8.1, which illustrates the discussion that follows in terms of this layered structure) is clearly not relevant only to measurement. For example, in introducing Shannon's theory of transmission, Warren Weaver pointed out (Shannon & Weaver, 1949: p. 4) that

> relative to the broad subject of communication, there seem to be problems at three levels. Thus it seems reasonable to ask, serially:
> Level A. How accurately can the symbols of communication be transmitted? (The technical problem.)
> Level B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
> Level C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.).

Despite some lexical peculiarities, Weaver's analysis is helpful for understanding the framework of semiotics.

A) What Weaver calls *the technical problem* only requires that a set of elements is given, possibly together with some conditions of comparison and rules of combination among such elements. Despite its simplicity, this standpoint proved to be very effective in grounding Shannon's mathematical theory of transmission, and from then on in providing a criterion for measuring the quantity of information in bits (i.e., if the set contains two equiprobable elements, the selection of one of them conveys 1 bit of information, and so on). Within this layer, information is called *syntactic*, being about "the formal relations of signs to one another", in the words of Charles Morris (1946: p. 217). We will use "data" and "syntactic information" as synonyms here. Data as such are purely formal entities, whose treatment – storage, retrieval, transfer, and also processing whenever the rules of combinations are explicitly specified – does not depend on the meaning(s) that someone might associate with them. Data acquisition is then where the process is grounded. But this still leaves open the question of what data is. Let us rely again on Weaver's presentation: "To be sure, [data] relates not so much to what you *do* say, as to what you *could* say. That is, [data affects] one's freedom of choice when one selects a message. If one is confronted with a very elementary situation where he has to choose one of two alternative messages, then it is arbitrarily said that the [syntactic] information, associated with this situation, is unity. [...] The two messages between which one must choose, in such a selection, can be anything one likes." (p. 9).[1] This leads to the basic understanding of data as support of selection of difference – it is this, but it might have been that. That is why any binary system is structurally the simplest provider of data: it is 0 (or false, or white, or...) but it might have been 1 (or true, or black, or...). More generally, whenever a set of possibilities, $X = \{x_i\}$, is available, then the selection of one or more of its elements is a provider of data – i.e., it is what is commonly referred to as "the raw data".

B) What Weaver calls *the semantic problem* broadens the scope of the technical / syntactic problem, by acknowledging that we usually acquire and manage data for referring to something, not to perform a purely syntactic activity. Despite the wealth of results that can be obtained at the syntactic layer – as paved by Shannon's two fundamental theorems about source entropy and channel capacity – our interest is usually focused on data as carriers of

[1] This analogy between communication and measurement should be read while keeping attention also to their substantial differences, as highlighted in Sect. 4.2.1.

meanings. In other words, data may have a meaning, if the elements of the set refer to, or stand for,[2] something else outside the set itself. This merges them into a second layer, in which the emphasis is on "the relations of signs to the objects to which the signs are applicable" (Morris, 1946: p. 217). Data equipped with meaning is called *semantic* information. The core arguments of this book relate to this: information about the measurand.

C) Finally, what Weaver calls the *effectiveness problem* builds upon the semantic layer and adds the context in which data-with-meaning is used by some agents for some purposes, where fitness for purpose is sometimes called the "value" of data.[3] Indeed, the same syntactic entity, e.g., the string "n-o", once equipped with a meaning, e.g., negation in English, and thus made a semantic entity, may have very different values (compare receiving a "no" to "have you already read my draft?" and "is all quiet on the western front?": quite a critical difference for most persons, though the syntax and semantics are exactly the same). Such an encompassing perspective is then about "the relations of signs to the interpreters" (Morris, 1946: p. 217), and it refers to what is called *pragmatic* information. From the perspective of this book, this encompasses questions about how measurements of the measurand can be used for particular purposes.

The fact that this is a layered structure explains why "any limitations discovered in the theory at Level A necessarily apply to levels B and C [... and] the analysis at Level A discloses that this level overlaps the other levels more than one could possibly naively suspect" (Shannon & Weaver, 1949: p. 6). And similarly, one could argue about how limitations at Level B apply to Level C. Along the same lines but in more explicit reference to language, Rudolf Carnap noted that "if in an investigation explicit reference is made to the speaker, or, to put it in more general terms, to the user of a language, then we assign it to the field of pragmatics. If we abstract from the user of the language and analyze only the expressions and their designata, we are in the field of semantics. And if, finally, we abstract from the designata also and analyze only the relations between the expressions, we are in (logical) syntax." (1942: p. 9).[4]

The parallel is then manifest:

- syntax is about data, i.e., signs that are distinguishable and that *may stand for* something, though this relation is still not included in the analysis;
- semantics is about data *equipped with meaning* due to its being related to objects, and
- pragmatics is about information *in a context* due to its being related to persons with purposes.

Three strategic clarifications are in order. First, this framework does not require us to accept that semantic information is always based on data or that pragmatic information is always based on semantic information, but only that these (possible) restrictions apply, in particular, to measurement-

[2] The relation *sign stands for entity* is very general. Famously, Charles Sanders Peirce identified three ways in which it can be realized. "If we come to interpret a sign as standing for its object in virtue of some shared quality, then the sign is an *icon*. Peirce's early examples of icons are portraits [...]. If [...] our interpretation comes in virtue of some brute, existential fact, causal connections say, then the sign is an *index*. Early examples include the weathercock, and the relationship between the murderer and his victim [...]. And finally, if we generate an interpretant in virtue of some observed general or conventional connection between sign and object, then the sign is a *symbol*. Early examples include the words 'homme' and 'man' sharing a reference." (Atkin, 2013; emphasis added). In this semiotic perspective, indication values (i.e., local values) can be interpreted as indexes of measurands, and measured values (i.e., public values) as icons of measurands.

[3] Of course, this is the concept of <value> related to goodness (Schroeder, 2016), which is different than <value of a property> of which we have been concerned throughout the entire book and explored particularly in Chap. 6.

[4] In the last few decades semiotics has had important developments (see, e.g., the overview by Wolf, nd) but – as with philosophy of language in Chap. 5 – a reference to some key founders and some of their main themes is sufficient for our purposes here. Its context is provided by the "semiotic triangle" that was introduced in Box 2.1.

related information. Second, information relevant to measurement is only descriptive, though possibly the basis for explanatory, predictive, or prescriptive purposes. Third, this framework does not imply that data is linguistic: what data is and how data can have or be provided with meaning are questions admitting multiple answers. This third point deserves a few more words.

In commenting on Peirce's definition of <sign> as "anything which is so determined by something else, called its Object, and so determines an effect upon a person, which effect I call its interpretant, that the latter is thereby mediately determined by the former", Albert Atkin (2013) is explicit that "we can think of the sign as the signifier, for example, a written word, an utterance, smoke as a sign for fire etc. The object, on the other hand, is best thought of as whatever is signified, for example, the object to which the written or uttered word attaches, or the fire signified by the smoke". Hence signs, and data, can be entities of a language – like English words – but can also be non-linguistic entities – like smoke – which

- are distinguishable (smoke vs. non-smoke: the syntactic layer),
- are possibly somehow connected to something else (the fire: the semantic layer), and
- are possibly relevant for decision making (call firefighters: the pragmatic layer).[5]

Analogously, data acquired by means of measurement is not linguistic; rather, it is represented by linguistic entities – typically, numerals and names of units in the case of quantitative properties – but it is not linguistic as such, as was shown in the discussion of values of properties in Chap. 6.

8.1.2 A semiotic perspective on measurement

An analysis of measurement as a process of (syntactic, semantic, and possibly pragmatic) information acquisition throws further light on the connections between the empirical and informational components of the process, which we now interpret according to the model-based realism proposed and justified in Sect. 4.5. In turn, this will allow us to finally identify what fundamentally characterizes measurement itself.

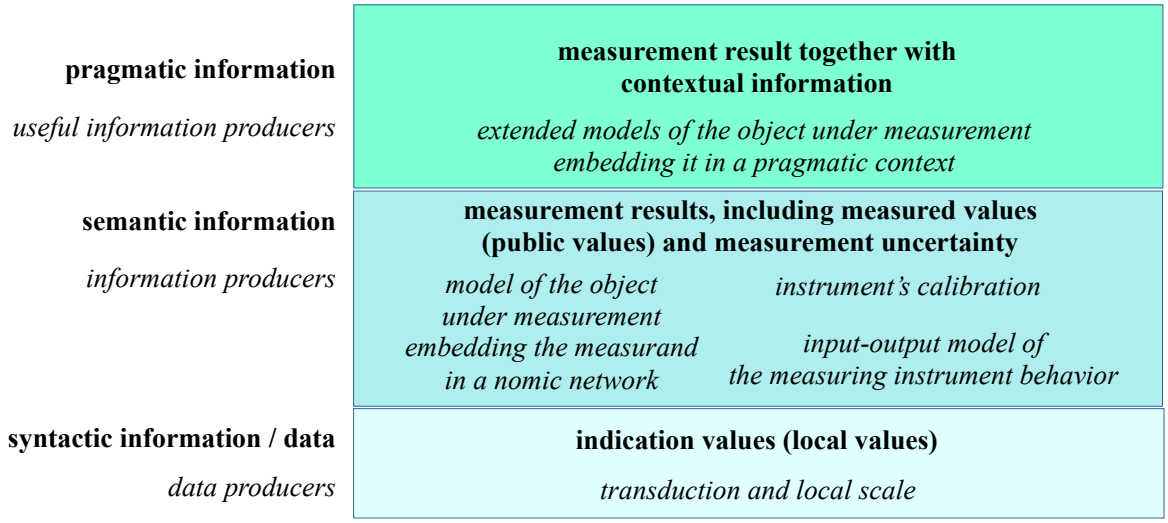


Fig. 8.1 The syntactic-semantic-pragmatic information layered structure in the perspective of measurement

[5] Since data can be non-linguistic, a further layer may be introduced for distinguishing between data and its representation. Depending on the context, smoke might be represented by the English word "smoke", the Italian word "fumo", etc. The fact that data, if it is not a linguistic entity, needs to be represented in order to be manipulated, communicated, etc., is a precondition for introducing the relations among syntactic, semantic, and pragmatic information. In what follows we do not deal with the problem of representation of data (an example is about the decimal representation of non-integer numbers: should the number 3/2 be represented as "1.5" or "1,5"?). The reader interested in notational issues may refer to the SI Brochure, Section 5, "Writing unit symbols and names, and expressing the values of quantities" (BIPM, 2019).

Let us consider the model of direct measurement presented in Sect. 7.3 from the perspective of the distinctions between and the relations among syntax, semantics, and pragmatics. Again, examples could be the measurement of temperature by means of an alcohol thermometer, which operates by transducing temperatures to positions of the upper surface of the alcohol in the instrument's tube, or the measurement of reading comprehension ability by means of a test, which operates by transducing abilities to patterns of responses.

As widely discussed in the previous pages and then formalized in the Hexagon Framework, the core empirical component of a measuring instrument is a transducer, that interacts with the effective property and produces a transduced property, i.e., an instrument indication, in response. With the assumption that the transduced property is observable or can be made observable, whether by human beings or technological devices, this input-output behavior is the bottom line of the process: if everything that generates the transduced property is kept within the black box and no assumptions are given about the input of the process (i.e., the effective property and the influence properties), then the transduced property, and therefore the indication value obtained via matching and scale application *is the measurement data*. Indeed, in this case the basic (syntactic) understanding of data as support for selection of differences applies, and in particular whenever a transducer invariably produces one and the same transduced property (i.e., its output is constant), thus in a condition of "no freedom of choice" in Weaver's words, we may safely conclude that such a device is unable to produce data. The local scale as such is in fact what measurement data is about: nothing more than that can be obtained by measurements performed by the instrument in this case, and entropy, in Shannon's sense, provides a means for its characterization.[6]

However, the assumed conditions make the system so simple that it is basically useless. Even the low-level features of measuring instruments introduced in Sect. 3.2.1, i.e., sensitivity, selectivity, and so on, cannot be evaluated if the only condition is the observability of the transduced property, and nothing is hypothesized about the cause of the observed effect. For example, the observation of smoke (as differentiated from non-smoke) as pure data is still not sufficient to infer the presence of fire.

Measurement data becomes semantic information only when at least a primitive version of both a model of the measurand and a model of the measuring instrument behavior is adopted (about these models and their relations see Sect. 7.2):

- the simplest version of a model of the measurand assumes that the property intended to be measured exists and the object under measurement carries an instance of it, with no further specifications;
- the simplest version of a model of the measuring instrument behavior assumes that the transducer is sensitive to the property intended to be measured and connects it causally to the transduced property, with no further specifications.

The fact that via these two models the property under measurement is embedded in a nomic network – though possibly a very simple one that connects the property under measurement only to the transduced property – shows that the adoption of these two models is definitely not a trivial move: rather, it is, conceptually and sometimes also operationally, a huge leap. Only on this basis does the

[6] For any given set $X = \{x_i\}$ equipped with a probability distribution such that $p(x_i)$ is the probability of selection of x_i (so that $\sum_i p(x_i) = 1$), the quantity of information – which should then more specifically called "quantity of syntactic information" or "quantity of data" – conveyed by the selection of x_i is $-\log(p(x_i))$. Accordingly, Shannon's entropy, $-\sum_i p(x_i) \log(p(x_i))$, can be interpreted as the average "amount of freedom" in the selection of elements from X. The maximum freedom is when the probability distribution is uniform, and therefore it does not add any constraints to the definition of the set; as mentioned above, the minimum freedom – zero entropy and in fact no freedom at all – is when one element is certain, and therefore all other ones are impossible. From the semiotic perspective we are discussing, this is a purely syntactic characterization: only data is involved, with no references to the property under measurement as its possible meaning.

transduced property convey semantic information on the property under measurement, and smoke becomes a sign of fire.

As discussed in Sect. 7.2.2, improving the model of the measurand involves distinguishing between the intended property and the effective property and identifying the role of affecting properties, and improving the model of the measuring instrument behavior involves distinguishing between the contributions of the effective property and the influence properties to the transduced property. These improvements make it possible to evaluate (a) the sensitivity of the transducer by assuming that the effective property may be changed in a controlled way while all influence properties are held constant, or, vice versa, (b) the selectivity of the transducer with respect to a given influence property by assuming that the influence property may be changed in a controlled way while the effective property and all other influence properties are held constant.

Hence a pre-measurement, as performed through a sequence

transduction → matching → local scale application

as introduced in Sect. 7.3.1, *produces the simplest case of measurement-related semantic information*: if two pre-measurements with the same measuring instrument produce distinct local values, and the instrument behaves sufficiently well (in terms of its sensitivity, selectivity, etc.) then we may infer that the two corresponding measured properties are distinct.

Nevertheless, the process cannot yet be considered a measurement (which is why the term "pre-measurement" was coined by Frigerio et al., 2010) for at least two reasons: the generated information

(i) is reported in terms of values in the local scale, which are values of the transduced property, i.e., the instrument indication, which is usually not of the same kind as the property under measurement, and

(ii) is also about the instrument, not only the object under measurement.

A next stage is then required, in which the metrological system offers a structural solution to these problems. Through instrument's calibration, local values are mapped to public values, which (i) are of the same kind as the property under measurement and (ii) provide information independent of the instrument. Even though we are not able to quantitatively evaluate this contribution of calibration,[7] it is clear that information is thus increased: accordingly, calibration enhances the model of the measuring instrument behavior, and makes it possible to report measurement results as Basic Evaluation Equations (where, by the way, the metaphor of smoke and fire no longer applies: the outcome of the inference is not that smoke and fire are equal) together with measurement uncertainty. Calibration *increases measurement-related semantic information*, and in fact makes measurement recognizable as such.

This semiotic perspective highlights the semantic role of measurement, but at the same time encompasses the pragmatic scenario in which measurement acquires a key role of enabler of information-enabled decision-making processes (Mari & Petri, 2017) through the comparison of measurement uncertainty with target uncertainty, the condition – as discussed in Sect. 7.4.4 – for considering measurement results actually useful to support the decision for which the measurement itself was designed and performed. In this broader context the evaluation of the quality of

[7] The foundational work made some decades ago for establishing a quantitative basis of (semantic) information – sometimes presented in terms of amount of *content* – did not lead to anything comparable to what Shannon's entropy constitutes for the quantitative evaluation of the amount of (syntactic) data (see, e.g., the extensive analysis in Hintikka, 1970). From this perspective it is unfortunate that the basic mathematical entity of Shannon's theory, $-\log(p(x_i))$, has been called "quantity of information" instead of "quantity of data". The usual remedy is to specify "quantity of syntactic information", or "quantity of statistical information", or also "quantity of technical information" in the lexicon adopted by Weaver, as mentioned above.

measurement and its results becomes a complex subject, in which, together with measurement uncertainty, several other conditions need to be taken into account, such as the timeliness of the acquired information and its pertinence to the decision to be made (Mari et al., 2012; Petri et al., 2021). The pragmatic import of measurement is thus well summarized: "There are things that can be measured. There are things that are worth measuring. But what can be measured is not always what is worth measuring; what gets measured may have no relationship to what we really want to know. The costs of measuring may be greater than the benefits. The things that get measured may draw effort away from the things we really care about." (Muller, 2018: p. 3).

Hence, measurement allows us to climb the semiotic layers, where, as depicted in Fig. 8.1, the sequence[8]

> data (syntactic information) → (semantic) information → useful (pragmatic) information

is then

> indication values → measurement results → measurement results in a decision-making context

Let us now review the main stages that have led us here and that, in the final section, will allow us to discuss the core question of this book: can there be one meaning of "measurement" across the sciences?

8.2 The path we have walked so far

In this book we have sought a characterization of <measurement> capable of explaining the acknowledged epistemic authority of measurement, but not needlessly tied to a specific subject matter or algebraic constraints.

Our starting point, in Chap. 2, was the identification of a basic set of necessary conditions for measurement, hypothesized to be plausibly acceptable by most, if not all, researchers and practitioners. The outcome of that chapter was the statement that

> measurement is an empirical and informational process, designed on purpose, whose input is an empirical property of an object and that produces information in the form of values of that property

Chap. 3 added three key specifications to this standpoint. First, though sometimes neglected in non-scientific situations, *measurement results should include information about the quality of the reported values*, which in the past was described in reference to measurement errors but is today more usually modeled in terms of uncertainty and validity, in physical and psychosocial measurement, respectively. Second, measured values report a relational form of information – the ratio of the measured property to the chosen unit, in usual quantitative cases – as inherited from the Euclidean tradition: this requires the social availability of a metrological system aimed at disseminating the reference properties by means of measurement standards mutually connected in traceability chains. Hence *measurement*

[8] Some characterizations of measurement require this entire sequence to be completed for a process to be actually considered a measurement. A significant example is the NIST definition, already quoted in Footnote 11 of Chap. 5, which includes the specification that measurement is a process "intended for use in support of decision-making" (Possolo, 2015: p. 12). Accordingly, any evaluation process only aimed at increasing the information available on a concerned property would not fulfill a necessary condition to be a measurement. Consider, for example, the observation and recording of angular positions of stars as performed be means of a telescope. Admittedly, the concept <decision> is generic, but either it is meant in so a generic sense that also in this case a decision is involved (perhaps the decision whether to add new data to the catalogue?), or this would not be a case of measurement, a paradoxical conclusion. Given this, and while acknowledging that the pragmatic layer provides an important application perspective, we think that measurement is an enabler of decision making, not that measurement must include a decision making stage.

requires calibration, and measurement standards make the calibration of measuring instruments possible. Third, despite its historical importance, the actual relevance of the Euclidean tradition to measurement science has been overemphasized: indeed, it refers to the mathematical concept <measure>, i.e., a number as a ratio of entities, which is only loosely related to the above-mentioned empirical and informational process of measurement. The point is then that *the condition that measurement applies only to quantitative properties cannot be justified by reference to the Euclidean tradition*.

The rest of the book can be interpreted as a report of our explorations around one question: *given these necessary conditions, what complementary conditions are sufficient to characterize measurement?*

With that in mind, we discussed, in Chap. 4, the epistemic status of measurement and the conditions of its proper use, as understood in the context of the three broad perspectives of realism, operationalism, and representationalism. The main findings were presented in a simple two-by-two matrix whose dimensions specify whether measurement has been characterized as being dependent on empirical and/or mathematical constraints, respectively, which led to the conclusion that what characterizes measurement is the empirical structure of the process, not some mathematical constraints on the inputs or the outputs of the process. This is in fact the position that we have developed, coupled with the acknowledgment of the unavoidability of the role of models in the process, thus grounded by what could be called a *model-dependent realism about measurement*.

Not surprisingly, the next stage of the exploration was about the very target of measurement, i.e., properties, which were analyzed in Chap. 5 from both ontological and epistemological perspectives. Here the core issue is as simple as it is controversial, in that it concerns the actual meaning of the Basic Evaluation Equation

property of a given object = value of a property

which is the basic structure of any measurement result, as complemented with information about measurement uncertainty. Consistent with our model-dependent realist standpoint, we interpreted this relation as *the claim of an actual referential equality*: it conveys information on the measurand because the measurand and the measured value are conceptually distinct entities, though they identify the same individual property. Given the conditions that measurement is an empirical process and that empirical processes cannot be performed on conceptual or mathematical entities, this forced us to take on an analysis of the existence of properties. The complexity of this subject is also due to the fact that <property> is a cluster concept, encompassing four sub-concepts: <property of an object> (e.g., the mass of a given object and the reading comprehension ability of a given individual), <value of a property> (e.g., 1.234 kg and 1.23 logits on a given RCA scale), <individual property> (e.g., a given mass and a given reading comprehension ability), and <general property> (e.g., mass and reading comprehension ability). From our model-dependent realist perspective individual properties exist as universals, but other positions are also compatible with the interpretation of the Basic Evaluation Equation as a referential equality, and thus possible disagreements over the actual nature of the entities exemplified by one or more of the sub-concepts of <property> did not block continued progress in our exploration.

Three fundamental issues for measurement science were then discussed in Chap. 6. The first was about the nature of values of quantities and more generally values of properties. A step-by-step construction was provided to show that *values are individual properties, identified as elements of a scale*, rather than symbols for the representation of properties. From this perspective, the difference between values of quantitative and non-quantitative properties is a matter of the structure of the scale

to which they belong. The second issue was then about the structure of scales and the related conditions of invariance, which provided a criterion for classifying property evaluations and then properties themselves in terms of scale types. This analysis found no unique condition for separating quantitative and non-quantitative properties, and reinforced the position that *being quantitative and being measurable are distinct conditions*. The third issue concerned the conditions of the existence of general properties and the possible role of measurement in the definition of general properties. Our basic assumption was that an empirical process can interact only with an empirically existing entity, and that this applies both to the objects that bear the properties and the properties of the objects. Thus, the distinction needs to be maintained between empirical properties and mathematical variables that may be used as models of properties. *The hypothesis of existence of an empirical property is corroborated by the observation of effects causally attributed to the property*.

In Chap. 7 we finally proposed a general model of a measurement process, consistent with the ontological and epistemological commitments developed in the previous chapters. The distinction between empirical and informational processes was, again, the starting point: measurement is neither a purely empirical nor a purely informational process. We broadly distinguished between direct and indirect methods of measurement as a fundamental classification of measurement methods related to the complementary roles of empirical and informational components, where each indirect measurement necessarily includes at least one direct measurement. As a consequence, *a structural characterization of direct measurement is the actual foundation of measurement science*: this is what we proposed with the Hexagon Framework, and exemplified in reference to cases of both physical and psychosocial properties. The Framework was also used to highlight once again the importance of evaluating the quality of the information produced by a measurement, now described in terms of the high-level, complementary requirements of object-relatedness ("objectivity") and subject-independence ("intersubjectivity"). Finally, the Framework provided a sufficient condition for measurability: a property is measurable if it is the input of at least one process that has been successfully structured according to the Framework.

On this basis we may come back to our opening question: given the necessary conditions discussed in the first stages of our exploration and the conclusions reached in the subsequent stages, what complementary sufficient conditions can we propose for characterizing measurement across the sciences?

8.3 Can there be one meaning of "measurement" across the sciences?

As we have already discussed in this book, the *International Vocabulary of Metrology* (VIM; JCGM, 2012), possibly the most authoritative existing source regarding fundamental concepts in measurement and their related terms, gives the following definition of <measurement>: "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity". Our arguments have pointed that this definition could be revised in three aspects:

- given the diversity of cases of what an experiment can be (e.g., thought experiments, mathematical experiments, etc.), the condition that measurement is "experimental" is too weak: as discussed in Sect. 2.2.1, a measurement should be constrained first of all to be an *empirical* process, so as to avoid the conclusion that purely computational processes can be considered measurements; complementarily, as introduced in Sect. 2.3 and developed in Chap. 7, measurement must include also an *informational* component, which produces its result in the form of values of the measurand;

- what characterizes a property as a quantity is not so clear, as discussed in Sect. 6.5, and, apart from the traditional view that superposes <measurement> and <measure>, which we have criticized in Sect. 3.4.2, nothing clearly requires a *property* to be quantitative in order to be measurable;
- finally, the concept <property> is too generic, given that it encompasses both general properties and individual properties, as we discussed in particular in Sect. 6.1; individual properties can be identified both as properties of objects and values of properties, as summarized in Sect. 5.1.3: hence a more explicit reference is appropriate to the condition that what is measured is an *empirical property of an object*.

Hence our VIM-like, updated definition of <measurement> could be something like:

> process of empirically and informationally obtaining one or more values that can reasonably be attributed to an empirical property of an object

This remains, in essence, a black-box characterization of measurement, thus defined in terms of its input – an empirical property of an object – and output – one or more values – with the only further condition that the process that produces the output from the input is both empirical and informational.

Read at face value, this definition would equally apply to processes such as guessing or a statement of a personal opinion (e.g., "I think it will be three degrees warmer tomorrow than it is today"; "I'm ten percent happier today than I was yesterday"), with the only condition being that some empirical activity was performed for obtaining the result. In order to avoid the conclusion that such processes are included in the set of possible measurements, the VIM's definition critically relies on the adverb "reasonably". However, reasonableness is not a formal condition, and it is plausible that different fields of knowledge and cultural contexts would accept different criteria for identifying what is reasonable. Rather, the condition of reasonableness could be considered a strategy to provide an encompassing definition, acceptable by the broadest set of researchers and practitioners, independently of their field of activity (indeed, who would refuse to accept that measurement should produce "reasonable" results?). But this condition is not sufficiently specific: it could be that, for example, some guesses about empirical properties could be considered reasonable, but nevertheless we might refuse to accept them as measurements. Hence the problem we explore here is, as the conclusion of this book: along the line of the VIM's definition, can there be one *criterion of reasonableness* that is sufficient to characterize measurement processes across the sciences?

This key problem could be interpreted as related to a descriptive – rather than normative – characterization: "measurement" is a term for a designed, not a natural, entity, and in principle everyone is free to use the term for whatever (s)he likes. But this does not take into account the special epistemic authority commonly afforded to measurement: were it the case that measurement and processes such as opinion making were equally valuable, why should resources be devoted to developing, buying, maintaining, and operating expensive measuring instruments? Furthermore, we do not see, a priori, any reason for reasonableness to be restricted to the evaluation of physical properties: it is then a critical condition for the criterion we are seeking that it be one that applies *across the sciences*. A specific sufficient condition could be, in particular, that the process is realized by means of a properly operated calibrated physical transducer that is sensitive to the property intended to be measured, which is indeed the sufficient condition usually applied to characterize physical measurements, and, in fact, this prompted the initial development of the Hexagon Framework: it works for physical properties, and, in our view, also provides some guidelines for designing measurement systems for psychosocial properties. But treating this as a general sufficient condition would correspond to an old-fashioned physicalism, which assumes that only physical properties are

measurable, thus considering non-physical properties to be unmeasurable simply *by fiat*. Hence any such candidate criterion should be independent of domain-specific strategies of design of measuring instruments and implementation of measurement processes.

8.3.1 Different subject matters, different processes...

In the physical sciences, the properties intended to be measured are often already embedded in a well-established body of knowledge – a nomic network, as discussed in Sect. 4.3 – and the actual work involved in the design and operation of measuring instruments is usually aimed at the refinement of existing instruments or mathematical techniques. The term "instrumentation and measurement" (as found, for example, in the title of the IEEE *Instrumentation and Measurement Society* and *Transactions on Instrumentation and Measurement* journal) clearly conveys this message, and the books on measurement systems in these fields are largely devoted to presentation of the features and modes of operation of measuring instruments (e.g., Bentley, 2005; Doebelin, 1966). In the human sciences, analogously developed nomic networks do not exist, and thus activities aimed at measurement necessitate first of all the task of formally characterizing the properties of interest, in such a way that information on them can be acquired and then mathematically processed. The key issue in these cases is the validity of the hypotheses upon which the process of evaluation is based, which is the focus of measurement sub-disciplines within specific human sciences such as patient reported outcomes, educational and psychological assessments, and performance measurements in management.

This multiplicity of perspectives reflects the fact that researchers operating in different fields face quite different problems and challenges, and will thus have different associations with the concept of measurement. However, this does not preclude the possibility that there are common elements of the processes referred to as "measurements" in different fields, and that identifying these common elements may suggest paths towards the formulation of a conception of measurement applicable across the sciences.

8.3.2 ... with some structural commonalities...

A background commonality that has emerged in the development of this book[9] is that measurement (1) is both an empirical and an informational process that (2) produces information on empirical properties of objects (where objects can be bodies, systems, phenomena, events, processes, individuals, organizations, etc.), with such information being (3) in the form of values of the property intended to be measured. This is consistent with the VIM's definition quoted previously and, as with the VIM's definition, it is an abstract, black-box (i.e., input-output) characterization of measurement, which nevertheless serves to delimit its scope, in particular by distinguishing measurement from computation (where in fact the latter produces information from other information).

Even without "opening the box", such a characterization already includes several key conditions, related to

(i) *the procedural nature of the process*, highlighting that measurement is designed on purpose and therefore justifying the focus on the components of the process that implement the procedure,

[9] And from previous papers of ours, in which we have discussed the very definition of <measurement> (e.g., Mari, 2013), and its epistemology (e.g., Mari, 2003), the stereotypes that surround measurement (e.g., Mari et al., 2017), and in particular the mistaken assumption that measurement is identical to quantification (Mari et al., 2017b).

(ii) *the definition of the property intended to be measured and to which the produced information is attributed*, highlighting that measurement requires modeling stages that are preliminary to data acquisition, and therefore

(iii) *the unavoidable compresence* in the process *of both empirical components*, which make the interaction with an empirical property possible, *and informational components*, which make the attribution of an informational entity possible.

As discussed throughout this book, these are *necessary* conditions for a process of evaluation of an empirical property to be considered a measurement. But, as stated, we are also looking for *sufficient* conditions.

8.3.3 ... and a common emphasis on trustworthiness...

As was previously observed, measurement is usually associated with the documented quality of its results, and therefore its *trustworthiness*. In Sect. 7.4.3 and elsewhere (e.g., Mari et al., 2012) we have argued that the trustworthiness (or "dependability"; see, e.g., Maul et al., 2018) of measurement results requires that they convey information that is object-related and subject-independent, or, more explicitly:

(1) information that is specific to the measurand and, therefore, to a given property of the object under measurement. This means that the provided information should be independent of any other property of the object or the surrounding environment, which includes both the measuring instrument and the subject who is measuring. This corresponds to guaranteeing that measurement results actually provide information about the measurand and not some other property. This condition is about the appropriate attribution of information to its claimed object: hence, it is a requirement of object-relatedness, or *objectivity* for short;

(2) information interpretable in the same way by different users in different places and times. This corresponds to guaranteeing that measurement results are expressed in a form independent of the specific context and only referring to entities which are socially accessible, so that the meaning of a measurement result is unambiguous and can be easily reconstructed in principle by anyone, possibly on the basis of suitable conventions: hence, it is a requirement of subject-independence, or *intersubjectivity* for short.

According to this characterization, objectivity and intersubjectivity are independent features – something can be objective but not intersubjective, and vice versa – that identify the two dimensions of measurement: the claim of the possibility of obtaining information about empirical properties, and the claim of the possibility of socially reporting such information. It is thus through their objectivity and intersubjectivity that measurement results are considered trustworthy.

On the other hand, as intended here, objectivity and intersubjectivity are not Boolean (i.e., yes-no) conditions: something can be more or less objective and more or less intersubjective. Hence for objectivity and intersubjectivity to be the sufficient conditions we are seeking to characterize measurement as a specific kind of evaluation, a threshold of minimum objectivity and intersubjectivity should be set. Thus the term "reasonable" in the VIM's definition of <measurement> once again does the heavy lifting: a given attribution of values would be considered reasonable, and therefore, would be considered a measurement, if both its objectivity and intersubjectivity were *sufficient for the intended purpose* for which the instrument was designed and performed. This highlights the pragmatic nature of measurement: a given measurement may be considered good for some purposes and bad for some others.

Emphasis on sufficient objectivity and intersubjectivity for a given purpose is then operationally useful, for the general guidelines it provides regarding the design and performance of measurements (e.g., in Petri et al., 2015) – though more work should be done to develop better guidelines – but it is still too specific at least in one respect: it would assume that measurement is always *good* measurement. While pragmatically this is sound – we would avoid knowingly performing a bad measurement (e.g., whose measurement uncertainty is greater than target uncertainty) – as discussed in Sect. 7.4.4 the concept <bad measurement> as such is not contradictory, and bad measurements do not fulfill the condition of sufficient objectivity and intersubjectivity. In other words, in order to maintain the VIM's characterization of reasonableness, objectivity and intersubjectivity are useful but still not sufficient: some other condition has to be identified.

8.3.4 ... and a focus on producing explicitly justifiable information

Science aims at the development of knowledge, where, following Plato, knowledge is commonly understood as *justified true belief*.[10] But although science aims at truth, it cannot guarantee it, as plainly illustrated by its history, and thus the truth or falsity (and more generally the quality) of any given scientific theory cannot be a definitional component of what makes it scientific: a theory can be of low quality, and even eventually admitted to be false, and nevertheless can be scientific. Although discussions of the definition and essential features of science are still ongoing, it is generally agreed that scientific theories must be explicitly justifiable, in that their logical and evidentiary grounding must be clear and publicly evaluable (see, e.g., Hansson, 2017). Thus, science is characterized by its structure rather than its outcomes; it is no contradiction to say that a theory is both scientific and false, but it would be a contradiction to say that a theory is both scientific and untestable, even if the theory were true.[11]

While approaching the end of the path we have followed in this book, we state our belief that this feature of science applies equally well to measurement, and is in fact the sufficient condition we were seeking, for complementing the necessary conditions introduced in Chap. 2: the trustworthiness of measurement results is not earned solely by their objectivity and intersubjectivity, but first of all *by their justification*. More explicitly, the result of the evaluation of a property can be claimed to be a measurement result only on the condition that in principle it is possible to explain how it was obtained with sufficient clarity to allow its critical analysis by all relevant stakeholders.[12] What was mentioned

[10] "According to this account, the three conditions – truth, belief, and justification – are individually necessary and jointly sufficient for knowledge of facts." (Steup & Ram, 2020: 2.3).

[11] A clear and simple example of this fundamental characterization of science is given, by difference, by Daniel Dennett: "There are many strategies, some good, some bad. Here is a strategy, for instance, for predicting the future behavior of a person: determine the date and hour of the person's birth and then feed this modest datum into one or another astrological algorithm for generating predictions of the person's prospects. This strategy is deplorably popular. Its popularity is deplorable only because we have such good reasons for believing that it does not work. When astrological predictions come true this is sheer luck, or the result of such vagueness or ambiguity in the prophecy that almost any eventuality can be construed to confirm it. But suppose the astrological strategy did in fact work well on some people. We could call those people astrological systems – systems whose behavior was, as a matter of fact, predictable by the astrological strategy. If there were such people, such astrological systems, we would be more interested than most of us in fact are in how the astrological strategy works – that is, we would be interested in the rules, principles, or methods of astrology. We could find out how the strategy works by asking astrologers, reading their books, and observing them in action. But we would also be curious about why it worked. We might find that astrologers had no useful opinions about this latter question – they either had no theory of why it worked or their theories were pure hokum. Having a good strategy is one thing; knowing why it works is another." (1987: p. 16). We claim exactly the same of measurement: that its results work (in some sense) is not enough; we want to know *why* they work. And this requires "opening the box" of the process and examining its structure and functioning.

[12] A more specific condition is then that measurement results *need to be reproducible* by all relevant social stakeholders. Even neglecting the practical constraints related to the fact that setting up a measurement system may have costs which are not affordable for all interested parties, we must acknowledge that some measurements are not repeatable, for

above about science can be repeated about measurement, then: it is no contradiction to say that a given process is both a measurement and produces results that are of low quality (i.e., it is a bad measurement) but it would be a contradiction to say that a process is both a measurement and cannot be justified, even if its results were (perhaps accidentally) accurate. Of course, in many cases, particularly of non-scientific measurements, measurement systems remain black boxes and no justifications of their results are reported: what is required is that a justification *can* be provided, whenever required.

On this basis, we are finally ready to propose a characterization of <measurement> that includes both necessary and sufficient conditions:

> measurement is an empirical and informational process that is designed on purpose, whose input is an empirical property of an object, and that produces explicitly justifiable information in the form of values of that property

In this way, the expectations of good measurement processes are continuous with the expectations of other sources of knowledge, in Plato's sense of knowledge as justified true belief. This is the case regardless of the subject matter and the details of the procedures involved in measurement.

8.3.5 Consequences for the theory and the practice of measurement

The requirement that measurement results be explicitly justifiable helps explain why measurement cannot be adequately characterized solely using a black-box model: if a given attribution of value(s) to a property is claimed to be a measurement (instead of, e.g., once again, a guess), it must be possible to explain how it was performed, and this requires opening the box and identifying the features of the process that secure the quality of the results. Given the diverse contexts in which measurements are applied, it is not surprising that what is found inside the (metaphorical, but sometimes also actual) box is also diverse.

We claim that there is a commonality in this diversity, however, and this commonality is structural: no matter how complex is the measurement system, (i) it is based on one or more direct measurements, as discussed in Sect. 7.2, and (ii) the structure of a direct measurement is based on the Hexagon Framework, introduced in Sect. 7.3. The fundamental point here is that *it is the structure itself of the process that provides the required epistemic justification*,[13] through the components of the Framework whose behavior explains how a Basic Evaluation Equation is obtained. In other words, the condition of epistemic justification that characterizes measurement is embedded in the Framework, and therefore is inherited by any process that is structured accordingly.

The evaluation of uncertainty in measurement has an important role in this, as clearly presented in the opening section of the *Guide to the expression of uncertainty in measurement* (GUM): "When reporting the result of a measurement of a property, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its trustworthiness. Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a specification or standard." (JCGM, 2008: 0.1; adapted, having substituted "physical quantity" with "property" and "reliability", a term with several other technical meanings, with

example when they alter the state of the object under measurement in an irreversible way (see, e.g., destructive testing, www.electropedia.org/iev/iev.nsf/display?openform&ievref=151-16-29). Hence reproducibility cannot be taken as a characterizing condition.

[13] We are referring here to the epistemic justification of measurement results, and therefore to the principled possibility of interpreting the information produced by measurement in a social context where it becomes shared knowledge. Higher level forms of justification are not only possible but usually desirable for measurement, and in particular pragmatic justification, aimed at showing that the measurement results deserve the resources used for obtaining them.

"trustworthiness"). In any non-trivial measurement there are indeed multiple sources of uncertainty, and collecting and combining them in an uncertainty budget (JCGM, 2012: def. 2.33) requires the box to be opened and the features of what is inside, i.e., the models of the object under measurement and of the measurement, explicated. The position of considering measurement to be a process that produces explicitly justifiable information is then coherent with the quoted principle of the GUM:

(i) any measurement result conveys a given quantity of information on the measurand, such that the greater the conveyed quantity of information the higher the assumed quality of the result (the GUM refers to this in terms of "quantitative indication of the quality of the result");

(ii) the justification to be provided relates in particular to the quality of the result, and therefore to the quantity of information conveyed through it;

(iii) the quality of the result can be specified in terms of measurement uncertainty, such that the greater the quality the less the uncertainty.

From the point of view of the user of a measurement result, the information about the measurement uncertainty in the result has then the critical role of being an effective substitute for actually opening the box, of course under the condition that the uncertainty is evaluated and reported in a honest way. The GUM offers a very clear proviso about the fact that "the evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement. The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value." (JCGM, 2008: 3.4.8). Under these conditions, a measurement result that explicitly reports uncertainty is a pledge, taken by the measurer, that the result is justified.

Given this, it is, of course, no simple matter to describe the details, inside the box, of how such justification should take place in any given instance of measurement, and it is here in which we find significant differences across disciplines. The key activities associated with designing and using measuring instruments in the physical and human sciences remain different, which could give the impression to a casual observer that there can be no shared meaning of <measurement> in such diverse fields. However, a closer look reveals a common condition – the possibility of explicit justification provided by a common structure of the process – and thus helps clarify how the reasonableness criterion of the VIM's definition of measurement is applicable across the sciences.

By establishing the same kind of characterization for science and measurement – both as endeavors which produce results that must be explicitly justifiable – the strategic role of measurement in society is explained: measurement is a tool by which the effective methods of science are adopted across the sciences and, indeed, beyond science into the day-to-day world.

Box 8.1 – Toward a manifesto for a widespread metrological culture

Our society is complex: we are overloaded with data, but sometimes lack the competence to transform it into meaningful information and useful knowledge.

Facing our limited adequacy to cope with this data deluge, two opposing, extreme ideologies have been developed. On the one hand, the position sometimes referred to as *dataism* (Brooks, 2013) claims that, by exploiting suitably trained technological systems, big data is sufficient to appropriately make whatever decision we need: models of data producers and data production are no longer required. On the other hand, what might be termed a *post-truth ideology* (e.g., McIntyre, 2018) denies the very possibility of scientific evidence, by pushing the principle that there are no

facts, but only interpretations: everything is a model of something else.

The proven special efficacy of measurement hints at the possibility of a third, intermediate strategy for handling complexity, taking as its basis the scientific method as enabled by measurement science and the associated measurement technologies. Measurement – an empirical and informational process that is designed on purpose, whose input is an empirical property of an object, and that produces explicitly justifiable information in the form of values of that property – is instrumental for producing information that is sufficiently objective and intersubjective for its expected purposes, and therefore sufficiently trustworthy for use in decision making.

Indeed, the activity of measuring

- is aimed at producing information on empirical properties, provided in the form of values of these properties,
- operates with shared and previously accepted references, materialized in calibrated measurement standards, and guarantees the traceability of the results produced to such references, and
- works according to explicit and transparent procedures, so that, in principle, anyone can ascertain how its results are produced.

In different contexts and with different devices, everybody regularly performs measurements: making everyone aware of these features of what they are already doing and how these features, together, are the components of measurement, would spread a *metrological culture* helping to overcome the inherent limits of dataism, by recognizing that data is necessary but not sufficient, but without falling into the relativism of post-truth ideology, by recognizing that hypotheses concerning the empirical world must pass the scrutiny of empirical confirmation.

Our society needs mediators, people who can credibly stand in the middle, between opposite extremisms: *metrological culture should greatly contribute to this*.

References

Atkin, A. (2013). Peirce's Theory of Signs. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/peirce-semiotics

Bentley, J. P. (2005). *Principles of measurement systems.* Harlow: Pearson.

Brooks, D. (2013, Feb). Opinion – The Philosophy of data. *The New York Times*. ISSN 0362-4331.

Carnap, R. (1942). *Introduction to semantics*. Cambridge: Harvard University Press.

Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

Doebelin, E. (1966). *Measurement systems: Application and design* (5th ed. 2003). New York: McGraw-Hill.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*, 123–149.

Hansson, S. O. (2017). Science and Pseudo-Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/pseudo-science

Hintikka, J. (1970). On semantic information. In J. Hintikka & P. Suppes (Eds.), *Information and inference* (pp. 3–27). Dordrecht: Reidel.

International Bureau of Weights and Measures (BIPM) (2019). *The International System of Units (SI ("SI Brochure")* (9th ed). Sèvres: BIPM.

Joint Committee for Guides in Metrology (2008). *JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM)*. Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Joint Committee for Guides in Metrology (2012). *JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* (3rd ed.; 2008 version with minor corrections). Sèvres: JCGM. Retrieved from www.bipm.org/en/committees/jc/jcgm/publications

Mari, L. (2003). Epistemology of measurement. *Measurement, 34,* 17–30.

Mari, L. (2013). A quest for the definition of measurement. *Measurement, 46*, 2889–2895.

Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy of Science, 65–66*, 46–56.

Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: A pragmatic view. *IEEE Transactions on Instrumentation and Measurement, 61*, 2107–2115.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement, 100*, 115–12.

Mari, L., & Petri, D. (2017). The metrological culture in the context of Big Data: Managing data-driven decision confidence. *IEEE Instrumentation and Measurement Magazine, 20(5)*, 4–20.

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement, 116*, 611–620.

McIntyre, L. (2018). *Post-Truth*. Cambridge, MA: MIT Press.

Morris, C. (1946). *Signs, language, and behavior*. New York: Prentice-Hall.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton, NJ: Princeton University Press.

Petri, D., Mari, L., & Carbone, P. (2015). A structured methodology for measurement development. *IEEE Transactions on Instrumentation and Measurement, 64,* 2367–2379.

Petri, D., Carbone, P., & Mari, L. (2021). Quality of measurement information in decision making. *IEEE Transactions on Instrumentation and Measurement, 70*, 1–16.

Possolo, A. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST measurement results*. NIST Technical Note 1900. Retrieved from www.nist.gov/publications/simple-guide-evaluating-and-expressing-uncertainty-nist-measurement-results

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Schroeder, M. (2016). Value theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/value-theory

Steup, M., & Ram, N. (2020). Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. Retrieved from plato.stanford.edu/entries/epistemology

Wolf, M. P. (no date). Philosophy of language. In *The Internet Encyclopedia of Philosophy*. ISSN 2161-0002. Retrieved from www.iep.utm.edu/lang-phi

Appendix.
A basic concept system of measurement

The *International Vocabulary of Metrology* (VIM: JCGM, 2012) is the metrologist's reference document for basic and general concepts and associated terms about measurement and measurement systems. Whenever we have considered it possible, the analyses and discussions in this book draw from the VIM, by referring, often explicitly, to its definitions. However, despite its pivotal role, there are two reasons why the VIM is insufficient, and unfortunately sometimes inadequate, for a treatment of measurement across the sciences. One is that it only deals with measurement of physical properties (in a broad sense, thus including chemical properties, biological properties, etc.); the other is that the VIM has a strong focus on quantitative properties only. Both reasons justify the development of an even more fundamental and therefore encompassing concept system.

Furthermore, the VIM is a vocabulary, in the sense given by the International Organization for Standardization (ISO), i.e., a "terminological dictionary which contains designations and definitions from one or more specific subject fields" (ISO, 2000: 3.7.2), and as such it is bound for each defined concept to provide one definition, which is expected to be an intensional definition,[1] which is "preferable to other types of definitions and should be used whenever possible as [it] most clearly reveal[s] the characteristics of a concept within a concept system" (ISO, 2009: 6.2). This Appendix proposes a basic concept system of measurement with weaker terminological constraints, thus allowing us to provide, for each relevant concept, a term and a characterization, sometimes in the form of explanation rather than a definition,[2] thus summarizing what is presented and discussed in the book.

As is common in top-level / upper ontologies, we use the term "entity" to refer to the most generic concept, hence as a synonym of "object" in the sense of ISO 1087-1:2000: 3.1.1: "anything

[1] A structural strategy for building a concept system is top-down: some generic concepts are assumed without a definition – called "primitive concepts" or simply "primitives" – and other concepts are subsequently derived from them according to a conjunctive logic:

$X := Y_1$ and ... and Y_n

where the set $\{Y_i\}$ of the defining concepts is called the *intension* of the defined concept X. For example, the VIM definition of <measurement>, "process of experimentally obtaining one or more values that can reasonably be attributed to a quantity" (JCGM, 2012: 2.1), can be understood as a rephrasing of: measurement (X) is a process (Y_1) and (the process) is an experimental obtainment of values (Y_2) and is a reasonable attribution of (these) values to a quantity (Y_3), i.e., $X := Y_1$ and Y_2 and Y_3. Evidently, for such a definition to be well formulated the defining concepts (<process>, <experimental obtainment of values>, <reasonable attribution of values to a quantity>) must have been previously defined. In a concept system built according to this top-down strategy, definitions are means of specification: through definitions the system is built by progressive knowledge specification, where the relation between the defined concept X and each of the defining concepts Y_i is then *species-genus*, or, according to the ISO standards on terminology work, *subordinate-superordinate* (hence in the definition mentioned above <measurement> is a species / subordinate of the genus / superordinate <process>: measurement is a (species of / kind of) process). In an *intensional definition* (ISO, 2000: 3.3.2), one defining concept Y_1 is singled out as the superordinate, the remaining Y_2, ..., Y_n being its *delimiting characteristics* (ISO, 2000: 3.2.7). This leads to the template:

defined concept := *superordinate concept* such that *delimiting characteristics*

that can be read

X is a Y_1 such that Y_2 and ... and Y_n

so that, for example, a measurement (X) is a process (Y_1) such that it is an experimental obtainment of values (Y_2) and is a reasonable attribution of these values to a quantity (Y_3).

[2] In particular, the explanations in this concept system allow some circularities, i.e., the explanation of the concept X includes a reference to the concept Y, and the explanation of Y includes a reference to X. The substitution principle forbids this in a definition (ISO, 2009: 6.3.4). In other words, these explanations include both a(n informal) definition and some possible notes.

perceivable or conceivable", thus making it possible to use "object" for any entity that carries a property (hence in this sense properties are entities but are not objects). Together with "entity", many other, non-measurement-specific terms are used here with their usual meaning.

The concept system we propose is organized as a list of entries, where each entry is devoted to a concept and is organized as

- a sequential identifier,
- a list of one or more terms (whenever an entry contains two or more terms, they are considered to be synonyms),
- in parentheses, the reference to the section(s) of the book where the concept is introduced,
- the characterization (*not* a formal definition) of the concept, in which the first occurrence of a term whose concept is characterized in this concept system is underlined, and
- a short note about the relation between the proposed characterization and the corresponding entry, if any, in the VIM, where the identifier of the relevant VIM entry is delimited by square brackets.

Since the entries are listed in a conceptual top-down order, their alphabetical list is also provided here.

***** Alphabetical list of the entries *****

***

1. **object** (2.2): Together with property, this is such a fundamental concept that we only propose their inter-characterization: an object is an entity that has properties. Hence physical bodies, events, phenomena, processes, individuals, organizations, etc. are all examples of objects. Sets of objects are considered to be objects in turn.
Any object is characterized as distinct and identified with respect to a context / environment.
[the VIM takes this concept as primitive, and sometimes uses the phrase "phenomenon, body, or substance" (e.g., in [1.1]) for it]

2. **property** (2.1): Together with object, this is such a fundamental concept that we only propose their inter-characterization: a property is an entity that an object may have, and that is associated with a mode of empirical interaction of the object with its context / environment (hence, we are actually considering here empirical properties only). A relation that applies to a set of objects is a property of that set, considered as an object. For example, a relation of order is a property that pairs of objects may have.
A basic ontic distinction is between general properties and individual properties, such as length and any given length respectively.[3]
[the VIM takes this concept as primitive]
Fig. A.1 summarizes the key entities related to <property> and their relations, as interpreted in this book.

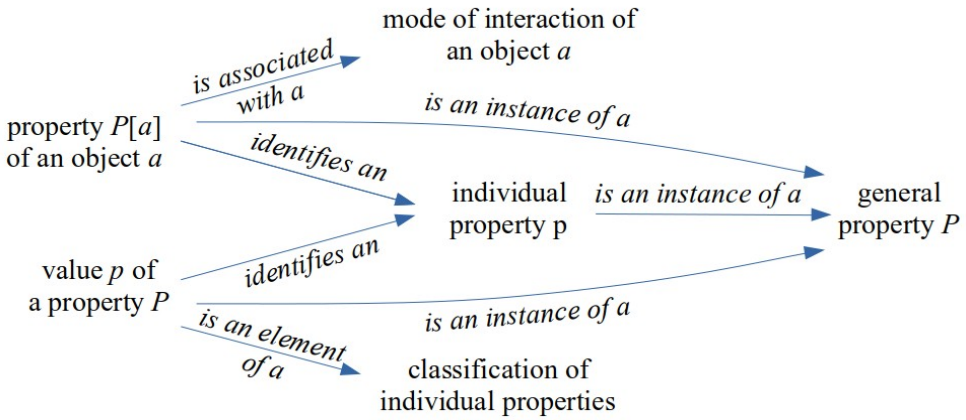


Fig. A.1 The key entities related to <property> and their relations

3. **general property, property in the general sense, kind of property, property** (2.2): An entity such as length, mass, shape, reading comprehension ability, socio-economic status, perceived quality, and so on. Any property under measurement is an individual property that is an instance of a general

[3] A justification of the choice of the adjectives "general" vs. "individual" as applied to properties is as follows:
– <individual> is the opposite of <general>, and in our case applies to properties (e.g., shape is a general property and a given shape, such as the shape of a given object, is an individual property);
– <specific> is the opposite of <generic>, and applies to concepts (e.g., <quantity> is more specific than <property>, in the sense that all quantities are properties, but there are properties that are not quantities);
– <particular> is opposite to <universal>, in our case referring to the ontological issue of whether an individual property cannot or can be shared by different objects, as discussed in Chap. 5;
– <concrete> is opposite to <abstract>, an even more complex and controversial distinction that we avoid here.

property. Two individual properties that are instances of the same general property are said to be "of the same kind".
There are several complementary ways to classify properties. The most important here is between empirical properties and non-empirical (e.g., informational, and more specifically mathematical) properties, given our hypothesis that only empirical properties can be measurable. Properties are classified by domain, e.g., physical properties vs. psychosocial properties. Properties, and more specifically their evaluations, are also classified by their types, where the simplest distinction is between quantitative and non-quantitative properties and evaluations. We do not emphasize here the traditional distinction between primary and secondary properties.
Entities such as systems of quantities and scales of properties refer to general properties.
In formal contexts (e.g., formulas) we write general properties with italic characters. Hence, *L* and *length* designate the general property length and *RCA* and *reading comprehension ability* designate the general property reading comprehension ability.
[the VIM does not define this concept; in the specific case of quantities, it uses "quantity" [1.1] and also "kind of quantity" [1.2]]

4. **individual property, property** (2.2): An entity such as a given length, a given reading comprehension ability, and so on. Any property under measurement is an individual property and therefore is an instance of a general property. An individual property can be identified as the property of an object or the value of a property. It is the invariant structure of relations among instances of the same general property, i.e., individual properties, that establishes the type of that general property.
In formal contexts (e.g., formulas) we write individual properties with lowercase roman or script characters. Hence ℓ designates a given length and r designates a given reading comprehension ability.
[the VIM does not define this concept; in the specific case of quantities, it uses "quantity" [1.1] or "individual quantity"]

5. **property comparison** (2.2): An empirical process by which individual properties of the same kind, i.e., instances of the same general property, are compared. The simplest case of comparison leads to property-related indistinguishability / distinguishability, or similarity / dissimilarity. Measurement is a property evaluation based on direct or indirect property comparison.
[the VIM does not define this concept]

6. **property evaluation, evaluation** (2.2): A process aimed at attributing a value to a property of an object, as reported in a Basic Evaluation Equation. Measurement is a property evaluation.
[the VIM does not define this concept]

7. **property evaluation type** (6.5): A criterion for classifying property evaluations, based on the algebraic invariance of values obtained through comparison. The simplest, but not uniquely defined, classification is between quantitative and non-quantitative evaluations. Types are (partially) ordered by the algebraic constraints that characterize them. For example, ratio type is algebraically stronger than interval type, because ratio type adds the constraint of a "natural zero" to interval type.
[the VIM does not define this concept]

8. **property type** (6.5): A criterion for classifying general properties, based on the algebraically strongest known evaluation type for the property. For example, a property is considered of quantitative type, i.e., a quantity for short, if at least one quantitative evaluation is known for it.

[the VIM does not define this concept]

9. **quantity**, **quantitative property** (2.2): A property of sufficiently strong type. There is not a clear-cut criterion for distinguishing quantitative and non-quantitative properties. For example, while ratio and interval properties are commonly considered to be quantitative, the VIM considers ordinal properties to be quantitative, whereas in other contexts they are sometimes considered to be non-quantitative.
[analogous to the VIM definition [1.1]]

10. **system of quantities**, **nomic network** (4.3) A set of general quantities together with a set of relations connecting them. The International System of Quantities (ISQ) is the most well known example of a system of quantitative physical properties. The term "nomological network" is also sometimes used for this.
[a generalization of the VIM definition [1.3]]

11. **property of an object** (2.1): An individual property concretely identified as a property that a given object carries, such as length of a given rigid body, the reading comprehension ability of a given individual, and so on. Any property under measurement is a property of an object. The set of all relevant properties of an object is supposed to provide an adequate description of the object, sometimes called its "state".
In formal contexts (e.g., formulas) we write properties of objects as $P[a, c]$ where P is the general property of which the considered individual property is an instance, a is the considered object, and c is the context (possibly including the time instant) in reference to which the object has been identified. In most cases, the reference to the context can be omitted and properties of objects are written as $P[a]$. Hence, $L[a]$ and $length[a]$ designate the length of rigid body a and $RCA[a]$ and *reading comprehension ability*$[a]$ designate the reading comprehension ability of individual reader a.
[the VIM does not define this concept; in the specific case of quantities, it uses "quantity" [1.1] or "individual quantity"]

12. **empirical property** (2.2): A general property whose instances are identified as modes of empirical interaction of objects with their context / environment, where the identification happens under the conditions that (i) an object may empirically interact with its context / environment in multiple modes, and each mode of interaction is considered to correspond to one property of the object, and (ii) some objects are comparable with respect to some of their properties, and sometimes distinct objects are discovered to have empirically indistinguishable properties.
[the VIM does not define this concept]

13. **object under measurement** (2.2): an object having a property which is the target of a measurement.
[the VIM does not define this concept]

14. **property under measurement** (2.2): A property of an object that is an instance of the general property intended to be measured in a given measurement. The individual property intended to be measured, i.e., the measurand, and the individual property with which the measuring instrument interacts, i.e., the effective property, may not be the same: by acknowledging the ambiguity of the specification "to be under measurement" (and also "measured"), both are called "properties under

measurement".
[the VIM does not define this concept]

15. **measurand, intended property** (2.3): A property of an object intended to be measured, and to which the measured values resulting from the measurement are attributed. Lacking a distinct term, we also use "measurand" to refer to the general property of which such property of an object is an instance. In the general case, the measurand and the effective property are distinct individual properties.
[the VIM has a definition of this concept [2.3], but only for quantities]

16. **effective property** (2.3): A property of an object, of the same kind as the measurand, with which the measuring instrument actually interacts. In the general case, the effective property and the measurand are distinct individual properties.
[the VIM does not define this concept, and sometimes uses "quantity being measured"]

17. **reference property** (2.3): A properly identified individual property chosen for comparisons with properties of objects of the same kind. In the case of ratio quantities, the unit is an example of a reference property. For operational purposes, reference properties are chosen to be properties of reference objects.
[the VIM does not define this concept]

18. **reference object, measurement standard** (3.3.1): An object that carries a reference property, associated with a value and usually an uncertainty, typically used in traceability chains to calibrate other measurement standards or measuring instruments.
[a generalization of the VIM definition [5.1], which only applies to quantities]

19. **scale of properties** (2.3): An invertible function f from a set $\{p_i\}$ of distinguishable reference properties for a general property to a set $\{\lambda_i\}$ of property identifiers, such that each identifier corresponds to one and only one value. For example, the scale of length-in-metres is such that f(the length identified as the metre) = 1, where the identifier 1 corresponds to the value 1 m, and the traditional definition of Celsius scale, i.e., the scale of temperature-in-degrees-Celsius, is such that f(the temperature of the boiling point of water at 1 atm pressure) = 100 and f(the temperature of the freezing point of water at 1 atm pressure) = 0, where the identifiers 100 and 0 correspond to the values 100 °C and 0 °C respectively.
[the VIM does not define this concept]0 °C

20. **public scale of properties** (7.3): A scale of properties based on reference objects that are expected to be socially available for supporting metrological traceability.
[the VIM does not define this concept]

21. **local scale of properties** (7.3): A scale of properties based on reference objects that are embedded in a measuring instrument.
[the VIM does not define this concept]

22. **unit** (2.2.4): A reference property for a general quantity, chosen to identify any other individual quantity of the same kind as a (not necessarily integer) multiple of it. The traditional term

"measurement unit", or "unit of measurement", is misleading given that units are used also outside measurement: a better expanded term is instead "quantity unit", or possibly "property unit".
[analogous to the VIM definition [1.9]]

23. **value of a property** (2.2): An individual property abstractly identified as an element of a given classification of individual properties of the same kind, and as such corresponding to one identifier in a scale of properties. Hence a value of a property is defined as the informational counterpart of a reference property in a scale. The specific mode of identification depends on the property type. In particular, values of quantitative properties are (not necessarily integer) multiples of an individual quantity chosen as the unit.
[the VIM has a definition of this concept [1.19], but only for quantities]

24. **value of a quantity** (2.2): An individual quantity identified as an element of a classification of quantitative properties, usually as the product of a number and a unit.
[analogous to the VIM definition [1.19]]

25. **numerical value of a quantity** (2.2.4): The number by which a unit is multiplied in a value of a quantity.
[analogous to the VIM definition [1.20]]

26. **Basic Evaluation Equation (BEE)** (2.2): An equation of the form

property of an object = value of a property

stating the equality of a property of an object and a value of a property, each asserted to be a conceptually distinct way of referentially identifying the same individual property. A Basic Evaluation Equation is the simplest case of a measurement result.
[the VIM does not define this concept]

27. **measurement** (Chapter 2, Chapter 7, and Section 8.3): An empirical and informational process designed on purpose, whose input is a property of an object and that produces justified information in the form of values of that property.
[a generalization of the VIM definition [2.1]]

28. **measurement method** (2.3): The structure of a process of measurement.
[analogous to the VIM definition [2.5]]

29. **direct (method of) measurement** (7.2.3): A method of measurement such that (i) a measuring instrument is used that is coupled with the object under measurement and is designed to interact with instances of the general property of the measurand, and (ii) the model of the measurand is only used in measurement for identifying the measurand.
[the VIM does not define this concept]
Fig. A.2 summarizes the black box model of the key entities related to <direct measurement> and their relations, as interpreted in this book (measurement uncertainty not included).

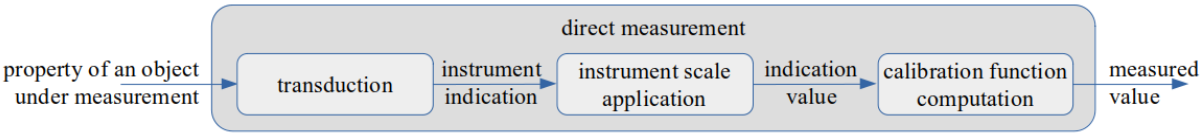


Fig. A.2 The black box model of the key entities related to <direct measurement> and their relations

Fig. A.3 summarizes the white box model of the key entities related to <direct measurement> and their relations, as interpreted in this book, i.e., the Hexagon Framework (measurement uncertainty not included).

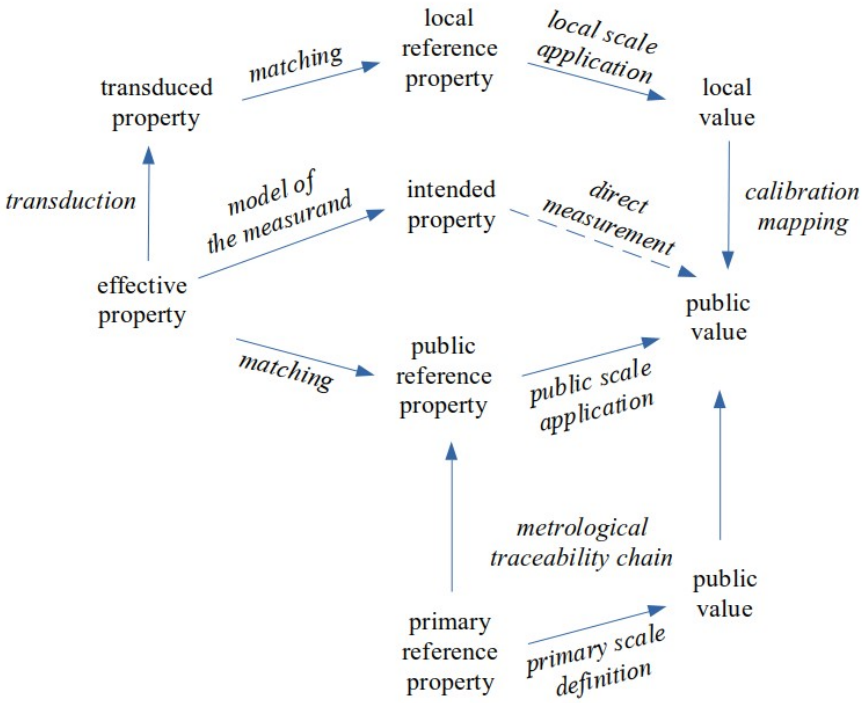


Fig. A.3 The white box model of the key entities related to <direct measurement> and their relations

30. **indirect (method of) measurement** (7.2.3): A method of measurement such that (i) one or more measuring instruments are used that are not necessarily coupled with the object under measurement or designed to interact with instances of the general property of the measurand, and (ii) the model of the measurand is used in measurement for identifying the measurand and its dependence on the properties from which the measurement result can be computed.
[the VIM does not define this concept]
Figure A.4 summarizes the key entities related to <indirect measurement> and their relations, as interpreted in this book (measurement uncertainty not included).

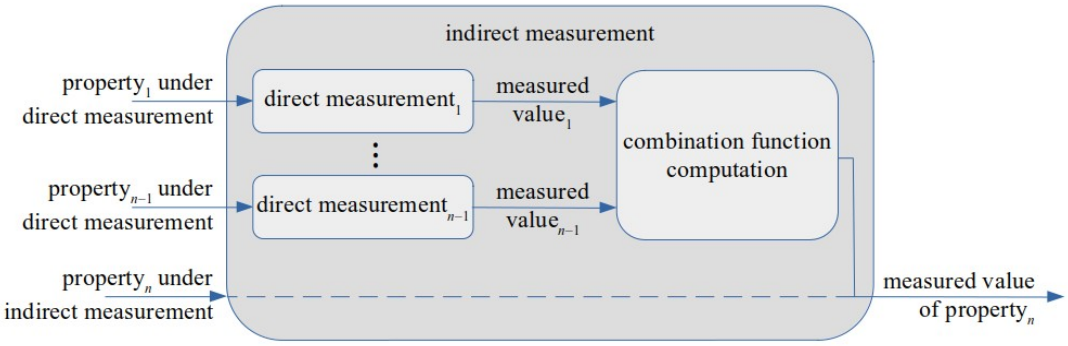


Fig. A.4 The key entities related to <indirect measurement> and their relations

31. **measurement procedure** (2.2.2): A specification about how a measurement is expected to be performed.
[analogous to the VIM definition [2.6]]

32. **measurement system** (2.2.2): A system, including empirical components, i.e., measuring instruments, and informational components, designed, built, set up, and operated so as to be able to interact with the instances of one or more general properties and to produce a measurement result as the outcome.
[the VIM does not define this concept]

33. **measuring instrument** (2.3): An empirical subsystem of a measurement system. We consider "measuring instrument" and "measuring system" as synonyms, and do not use the latter in order to avoid the ambiguity with "measurement system".
[analogous to the VIM definitions of <measuring instrument> [3.1] and <measuring system> [3.2]]

34. **measurement result** (2.2.4): The information produced by a measurement, in terms of values attributed to the measurand. While in the simplest cases it is a Basic Evaluation Equation, more complex options are used for reporting measurement uncertainty more explicitly, for example by substituting the measured value with a pair (measured value, standard measurement uncertainty), or an interval of values, or a probability distribution defined over the set of values.
Sometimes only the entity on the right-hand side of the equation, and therefore a measured value, is considered to be the measurement result.
[a generalization of the VIM definition [2.9]]

35. **measured value** (2.2.4): A value reported as a measurement result or as a component of a measurement result.
[analogous to the VIM definition [2.10]]

36. **measurement uncertainty** (3.2.3): An overall feature of the entities involved in a measurement, and inversely related to their quality: the traceability chain for the calibration of the measuring instrument (calibration uncertainty) and the behavior of the measuring instrument (instrumental uncertainty and interaction uncertainty).
Measurement uncertainty is part of any sufficiently well specified measurement result: in a well-designed measurement, measurement uncertainty is not less than definitional uncertainty and not greater than target uncertainty.

[analogous to the VIM definition [2.26]]
Fig. A.5 summarizes the key entities related to <measurement uncertainty> and their relations, as interpreted in this book.

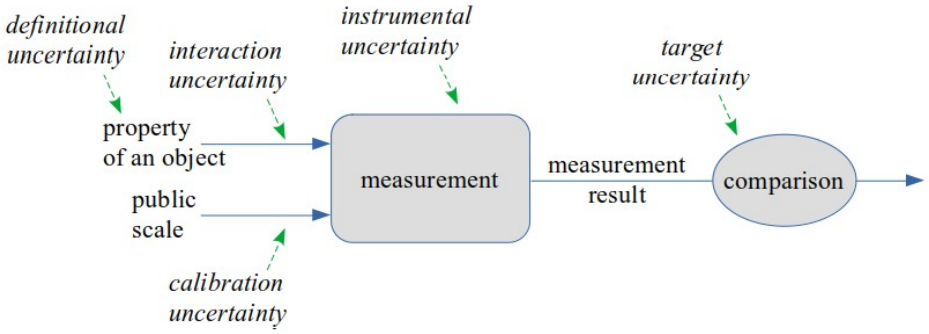


Fig. A.5 The entities related to <measurement uncertainty> and their relations

37. **definitional uncertainty** (3.2.4): An uncertainty about the measurand, as it is defined before its measurement is performed. Sometimes this cannot be expressed quantitatively. A measurement result whose uncertainty is less than definitional uncertainty corresponds to a waste of resources devoted to the design and performance of the measurement.
[analogous to the VIM definition [2.27]]

38. **calibration uncertainty** (3.2.4): A component of measurement uncertainty related to the public scale used for the calibration of a measuring instrument, and derived from the traceability chain that connects the scale to the realization of its definition.
[the VIM does not define this concept]

39. **interaction uncertainty** (3.2.4): A component of measurement uncertainty due to the fact that the interaction between the object under measurement and the measuring instrument can alter the state of the object itself.
[the VIM does not define this concept]

40. **instrumental uncertainty** (3.2.4): A component of measurement uncertainty due to the non-ideal behavior of the transducer in a measuring instrument, and in particular its limited selectivity and stability.
[analogous to the VIM definition [4.24]]

41. **target uncertainty** (7.3.4): An uncertainty specified as an upper limit, on the basis of the intended use of measurement results. A measurement result whose uncertainty is greater than target uncertainty is useless for its intended use.
[analogous to the VIM definition [2.34]]

42. **standard measurement uncertainty** (3.2.5): An uncertainty reported as a standard deviation of a probability distribution about the measurand.
[analogous to the VIM definition [2.30]]

43. **object-relatedness, objectivity of measurement** (7.4.3): An overall feature of measurement and its results, related to the extent to which the information conveyed by the measurement is about the

measurand and nothing else. Together with intersubjectivity, objectivity characterizes trustworthy measurements.
[the VIM does not define this concept]

44. **subject-independence, intersubjectivity of measurement** (7.4.3): An overall feature of measurement and its results, related to the extent to which the information conveyed by the measurement is interpretable in the same way by different persons in different places and times. Together with objectivity, intersubjectivity characterizes trustworthy measurements.
[the VIM does not define this concept]

45. **calibration** (2.3): An empirical and informational process performed on a measuring instrument for establishing a functional relation between the local scale of the instrument and a public scale, and therefore aimed at providing the traceability of measurement results obtained when using the instrument.
[analogous to the VIM definition [2.39]]

46. **metrological system** (3.3.1): A scientific, technological, and organizational system made of measurement standards and measuring instruments connected in traceability chains, aimed at guaranteeing the metrological traceability of the measurement results produced by such measuring instruments.
[the VIM does not define this concept]

47. **metrological traceability** (3.3.1): A feature of a measurement result of being related to a reference property through a documented unbroken traceability chain, each contributing to the calibration uncertainty.
[analogous to the VIM definition [2.41]]

48. **traceability chain** (3.3.1): A sequence of measurement standards terminated by a measuring instrument, all connected by calibrations and aimed at relating the measurement results produced by that instrument to the reference property borne by the first standard of the sequence.
[analogous to the VIM definition [2.42]]

49. **transducer, sensor** (2.3): A device that may be put in interaction with an object under measurement with respect to a property of the object, to which it is sensitive; as a result of the interaction, the device changes one of its properties, the instrument indication. A transducer is the core component of a measuring instrument.
(The concept <transducer> is in fact more general: both sensors and actuators are transducers.)
[analogous to the VIM definitions [3.7 and 3.8]]

50. **instrument indication** (2.3): A property of a measuring instrument, changes of which are causally dependent upon changes of the property to which the transducer of the instrument is sensitive, i.e., the effective property.
[analogous to the VIM definition [4.1]]

51. **affecting property** (3.2): A property of the object under measurement or its context, changes of which produce a change of the property under measurement.

[the VIM does not define this concept]

52. **influence property** (3.2): A property other than the property under measurement, changes of which produce a change in the behavior of the measuring instrument.
[analogous to the VIM definition [2.52]]

53. **transduction function** (7.2.3): A mathematical model of the behavior of the transducer in a measuring instrument, such that the indication is a function of the effective property and the influence properties.
[the VIM does not define this concept]

54. **calibration function** (7.2.3): A mathematical model reconstructing the behavior of a measuring instrument as the inverse of the transduction function.
[analogous to the VIM definitions [4.30 and 4.31]]

55. **correction function** (7.2.3): A mathematical model of the object under measurement as considered in a direct measurement, such that the measurand is a function of the effective property and the affecting properties. It may take into account bias.
[the VIM does not define this concept]

56. **combination function** (7.2.3): A mathematical model of the object under measurement as considered in an indirect measurement, such that the measurand is a function of other properties of the object or the context.
[the VIM does not define this concept]

57. **instrument accuracy** (3.2.1): An overall feature of a measuring instrument related to its ability to produce a measured value that is close to an accepted reference value, typically the value associated with the reference property of the measurement standard used to assess the metrological features of the instrument. It is similar to validity in the human sciences.
[analogous to the VIM definition [2.13]]

58. **instrument trueness** (3.2.1): A feature of a measuring instrument related to its ability to produce measured values whose average is close to an accepted reference value in a sufficiently large series of replicate independent measurements on the same or similar objects under specified conditions, where typically the value is associated with the reference property of the measurement standard used to assess the metrological features of the instrument.
[analogous to the VIM definition [2.14]]

59. **instrument precision** (3.2.1): A feature of a measuring instrument related to its ability to produce indication values or measured values close to each other in a series of replicate independent measurements on the same or similar objects under specified conditions. It is also known as reliability in the human sciences.
[analogous to the VIM definition [2.15]]

60. **instrument sensitivity** (3.2.1): A feature of a measuring instrument related to the ability of its transducer to produce changes of the indication in response to changes of the property under

measurement. Sensitivity is computed as the ratio of the change of indication values and the corresponding change of values of the property under measurement, while the influence properties are maintained constant.
[analogous to the VIM definition [4.12]]

61. **instrument selectivity** (3.2.1): A feature of a measuring instrument related to the ability of its transducer to not produce changes of the indication in response to changes of an influence property. Selectivity with respect to an influence property is computed as the inverse of the ratio of the change of indication values and the corresponding change of values of the influence property, while the property under measurement is maintained constant.
[a generalization of the VIM definition [4.13]]

62. **instrument stability** (3.2.1): A feature of a measuring instrument related to the ability of its transducer to not produce changes of the indication while time passes. Stability is computed as the inverse of the ratio of the change of indication values and the corresponding change of time instant values, while the property under measurement and the influence properties are maintained constant.
[analogous to the VIM definition [4.19]]

63. **instrument resolution** (3.2.1): A feature of a measuring instrument corresponding to the smallest change in the property under measurement that causes a perceptible change in the indication.
[analogous to the VIM definition [4.14]]

References

International Organization for Standardization (2000). *ISO 1087-1:2000, Terminology work – Vocabulary – Part 1: Theory and application*. Geneva: ISO.

International Organization for Standardization (2009). *ISO 704:2009, Terminology work – Principles and methods*. Geneva: ISO, 3rd ed.

JCGM (2012). *International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM)* 3rd ed (2008 version with minor corrections). Sèvres: Joint Committee for Guides in Metrology (www.bipm.org/en/publications/guides/vim.html).

Index of concepts and authors' names

(in square brackets the id of the entry in the concept system in Appendix; in parentheses the section(s) where the concept is referred to)

Index of figures

Index of tables

Index of boxes