preprint of a paper published in:Measurement, 116, 611-620, 2018, doi:10.1016/

The quality of measurement results in terms of the structural features of the measurement process

Andrew Maul[1#], Luca Mari[2$], David Torres Irribarra[3], Mark Wilson[4]

[1] Gevirtz Graduate School of Education, University of California, Santa Barbara, CA USA

[2] School of Industrial Engineering, Università Cattaneo – LIUC, Castellanza (VA), Italy

[3] Escuela de Psicología, Pontificia Universidad Católica de Chile

[4] Graduate School of Education, University of California, Berkeley, CA, USA

E-mail: amaul@ucsb.edu

Abstract. In both scientific and lay settings, measurement is considered a privileged source of high-quality information, and is commonly associated with precision, accuracy, and dependability. However, it is not always clear what features of the measurement process justify this public trust, and how the quality of measurement results in different domains of inquiry can be compared. In this paper, we first argue that the quality of measurement results depends on their object-relatedness ("objectivity") and subject-independence ("intersubjectivity") and is justified on the basis of the structural features of the measurement process, as well as features of the inputs or the outputs of the process. Given this perspective, we analyze three general measurement methods, according to which a measurement process can be structured and performed, which may be called (a) direct synchronous, (b) direct asynchronous, and (c) indirect. In addition to the value of these distinctions for the process of designing measuring instruments, they allow us to highlight the different roles of models, theories, and computations in measurement. We then attempt to apply this classification strategy in the context of the social sciences by discussing the role of (1) the definition of the measurand and (2) the theory connecting the measurand to the measurement results in each of these measurement methods, and how they can or cannot be conceptualized from the perspective of measurement theories in the social sciences. This leads us to the conclusion that the differences between physical and non-physical measurement are historical and contextual rather than essential; that is, in both cases, the quality of measurement results can be effectively evaluated from a structural perspective.

Keywords. measurement; philosophy of measurement; measure; quality of measurement

1. Introduction

It would be difficult to overstate the value and importance of measurement in nearly every aspect of society. Every time we eat food, take prescribed medicine, fly in an airplane, use a cell phone, 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 objectivity [Porter 2003]. Against this backdrop, it seems little wonder that the social sciences (including psychology, sociology, economics, and field-specific areas of research, such as education) 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 (see the review and related discussion in [Mari 2013]). In addition to obviously 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",

[#] To whom any correspondence should be addressed.

[$] One of the authors is a member of the Joint Committee on Guides in Metrology (JCGM) Working Group 2 (VIM). The opinion expressed in this paper does not necessarily represent the view of this Working Group.

measurement model", etc. – are used with very different and possibly even incompatible meanings, with potentially disastrous results.[1]

It would seem, then, that the clarification of foundational measurement concepts should (continue to) to be a high priority: in terms not only of the definition of measurement itself, but also of the identification of those features of measurement that justify its commonly-afforded degree of public trust and social prestige. Justification of the dependability of measurement results, in turn, depends on identifying those features of the measurement process that ensure (or, at least, confer high likelihood upon) the quality of the results.

There are at least two (categories of) reasons why measurement-related concepts have become so difficult to define in a consistent way across different fields. First, as the scope of measurement has broadened, it is not always obvious what – if indeed anything – is common among all the processes claimed to be measurements, but surely a shared body of knowledge cannot be found in the technical details on which measurement science advances in each specific field. Second, the scholarly treatment of the concept of measurement has focused since the second half of the 20th century on purely formal criteria, thus abstracting from the concrete realization of the process, up to the point that one of the reference books on representational theories of measurement is titled "Abstract measurement theory" [Narens 1985], and other researchers in the field have made 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] and "The theory of measurement is difficult enough without bringing in the theory of making measurements" [Kyburg 1984]. This emphasis on a formal characterization of measurement is consistent with the expansion of measurement into many new domains of application, abandoning definitions that could be tied to requirements of specific areas; abandoning for instance elements tied to the traditional realization of measuring systems operating on the basis of transductions implemented by physical sensors possibly due to the fact that the evaluation of non-physical properties[2] cannot conform to it. As a consequence, theoretical 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 instances of the expression of subjective judgments and opinions (as already acknowledged, e.g., by [Sawyer et al 2016]: "In the social sciences, in particular, most evaluations are not measure[ment]s, but rather mixtures of opinion and estimation.")

One may question whether working on the definition of 'measurement' is a worthwhile endeavor. Here our position on this matter is also practical: there is a social interest in sharing scientific and technical vocabulary across disciplines[3], particularly in the case of an infrastructural activity like measurement [JCGM 2012], 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 idea of "measurement" can be invoked at will, without understanding or concern for what has historically made it a valued practice, it becomes simply a rhetorical device, risking to discredit its practice in general.

[1] The comparability of measurement concepts and practices across different scientific disciplines has been the subject of a significant amount of scholarship over the past century. The present paper is aimed at contributing to the general endeavor of improving the understanding of measurement across the sciences, a complex subject on which one of the authors co-organized the 2016 Joint IMEKO TC1-TC7-TC13 Symposium, "Metrology Across the Sciences: Wishful Thinking?", 3–5 August 2016, Berkeley, USA, whose proceedings have been published in the IOP Journal of Physics: Conference Series, 772, 2016. Some other volumes that could be usefully considered on this matter are, e.g., [Berglund et al 2011], [Boumans 2015], [Boumans et al 2013], [Schlaudt & Huber 2015].

[2] For the sake of generality, the term "property" is used rather than "quantity" throughout this paper. The *International Vocabulary of Metrology* (VIM) defines quantities as specific kinds of properties [JCGM 2012, def.1.1].

[3] 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 electrical and electronic terminology database containing more than 20,000 terms and definitions", that makes the series of standards IEC 60050 freely accessible online at www.electropedia.org.

We propose here is that measurement is a process characterized by its *structure*, not only by the specification of the functional relationship connecting its inputs to its outputs: what is required is an explanation of *how* the process does what it does, not only of *what* it does. While, for example, measurements based on thermal expansion thermometers and on electrical resistance thermometers could be treated as interchangeable in functional terms, they clearly have different structures: even if the function is the same, its implementation / realization is distinct. Whereas a functional relationship relies solely on a black box model, a structural model involves identification of the invariant aspects that are implemented in the experimental process – and this, in turn, as we will argue, is what provides justification of the claim that measurement results are publicly trustworthy. As a corollary, any purely black-box (meta-)model cannot adequately account for relevant features of measurement, and thus is not sufficient for the purpose of understanding the quality of measurement results.

In the metrological tradition the general description of the structure of a measurement is provided by a so-called "measurement method", the "generic description of a logical organization of operations used in a measurement" according to the VIM [JCGM 2012, def.2.5]. This paper proposes some preliminary considerations and examples to show that different measurement methods, each with their own specific structures, share the same invariant meta-structure (on the concept of measurement meta-structure see also [Mari et al 2016], that the present paper expands). With some provisos – including the availability of a sufficiently well-detailed definition of the general property of which the measurand is an instance – this invariance is independent of the nature of the measurand and therefore spans the measurement of both physical and non-physical properties.

The next section is devoted to introducing this meta-structural understanding of measurement in reference to three basic measurement methods, as developed in metrology, and to discussing the conditions for the quality of measurement for each method. On this basis, section 3 explores how these structures apply in the case of non-physical properties, and argues that the most critical barrier to understanding the operative structure of non-physical measurement processes – and, therefore, to understanding how the dependability of such measurement results is justified – relates not to any fundamental distinction between the two areas, but to the often imprecise way in which general non-physical properties are defined.

2. A meta-structural understanding of measurement

2.1 Black-box characterizations of measurement

Under the general hypothesis that measurement is a process that operates on inputs (at least the measurand, in the case of direct measurement methods[4]) to produce outputs (at least the measurement result), measurement could be characterized as an instance of the black box meta-model that describes processes as entities that transform inputs to outputs (Pane [a] in Figure 1).

Conventionally, measurement is a process aimed at producing information in the form of values (e.g., 0.1234 m) attributed to properties (usually, quantities) of objects (e.g., the length of a given object). However, such a characterization is not specific to measurement: other processes, such as, say, quantitative guessing, take as input the property of an object and produce in output one or more values that are attributed to the property. Let us call "property evaluation", or simply "evaluation" for short, *any* process with this black box characterization (Pane [b] in Figure 1).

[4] The possibility of 'direct' measurement (i.e., more correctly, direct *methods* of measurement), is sometimes dismissed as naive, via the argument that "all measurements are indirect in one sense or another" because "not even simple physical measurements are direct" given that, e.g., "the physical weight of an object is customarily determined by watching a pointer on a scale. No one could truthfully say that he 'saw' the weight" [Guilford 1936]. Of course, this is not the meaning assumed here and, e.g., by the VIM [JCGM 2012, def. 2.5 Note], which notes that measurement methods are either direct or indirect. In this view, direct methods are simply those in which the measuring instrument directly interacts with the object under measurement.

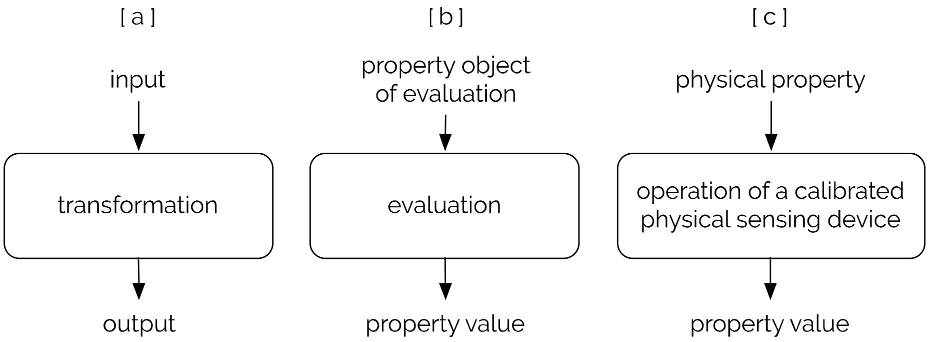

Figure 1. Three characterizations of physical measurement.

Hence measurements are evaluations but not all evaluations are measurements. This generates the question of identifying the conditions that make measurement a specific kind of evaluation. The interest of this issue is in the public trust acknowledged to measurement results, well grounded on the proven effectiveness of measurement in science, technology, health, trade, etc. and that is not equally shared by guessing and other kinds of evaluation: what is the source of this public trust? How is it justified?

A traditional position assumes that measurement is the unproblematic ground on which the scientific and technical development is based whenever reliable data is required, a "protocol of truth" in the classical terminology of philosophy [Margenau 1958]. This position holds that the observed variability of measurement results, due to the non-complete repeatability of measurement conditions, is accounted for in terms of measurement errors, hence assuming a purely empirical nature of the process. This standpoint, which systematically neglects a theoretical component of measurement, has been historically associated with the understanding of measurement as the evaluation of physical properties and a body of knowledge related to the realization of the generic input-output specification based on the adoption of physical instrumentation and standards. These two conditions correspond to the structural strategy of constraining both the input of the black box – only physical properties are measurable – and the contents of the black box (which then becomes a gray box) – only properly designed, set up, and operated physical sensing devices are measuring instruments. This is the traditional paradigm of physical measurement, in which measurement is intended as a process performed by a calibrated physical sensing device on an input physical property (Pane [c] in Figure 1).

Given the vast and growing multiplicity of different examples of (processes claimed to be) measurement, some characterizations have sought to define measurement, or locate necessary and/or sufficient conditions for measurement, in terms of features of the inputs and/or outputs of the process.

A first category of such views, such as those given by Bridgman [1927] and Dingle [1950], focuses on the outputs of the process. Dingle, for example, defines measurement as "any precisely specified operation that yields a number," thus referring to an aspect of the output of the process (i.e., that it is numerical) as a necessary and sufficient condition for measurement. In the absence of an account of what forms of precision are necessary, the requirement that the procedure be "precisely specified" seems to demand sufficient clarity regarding what the operations of the procedure are, but places no demands concerning the inputs of the procedure nor the manner in which they are transformed into outputs.

A second category of views focuses instead, or sometimes additionally, on the inputs of the process. For example, Michell (e.g., [2005]), in line with the Euclidean tradition, defines 'measurement' as the assessment of a quantity, the measurand, in comparison to a second quantity, the unit. According to this view, whether or not a given property is a quantity is an ontological issue, and is a pre-condition for measurement. In this view, a property that is not a quantity is a priori not measurable: thus a feature of the input, namely that it is a quantity, is a necessary condition for measurement [Mari et al. 2017]. Arguably, Michell's view also requires that the output of the process be numerical – specifically, that it be a ratio of a quantity value to a standard unit.

A third category of views characterizes measurement in terms of the formal relationship between inputs and outputs, but is silent regarding the structure of the transformation by which the relationship is implemented as a causal mechanism that induces variations in the outputs from variations in the inputs. Representational theories of measurement (e.g., [Scott & Suppes 1958], [Pfanzagl 1968], [Krantz et al 1971], [Roberts 1979], [Narens 1985, 2013, 2014]), for example, define 'measurement' in terms of a morphism between an empirical relational system and a numerical relational system; thus, in this view, a particular feature of the formal dependence of the output on the input is a (necessary and) sufficient condition for measurement.

As we have argued elsewhere (e.g., [Mari et al 2016], [Maul et al 2016]), while each of the views described above succeeds in capturing valuable intuitions about measurement, none of them provides a fully satisfactory set of conditions that could widely be used (a) to distinguish measurement from non-measurement processes, nor (b) to distinguish better (or more trustworthy, useful, etc) from worse instances of measurement. Views in the first category (focusing on outputs) tend to fail to disallow instances of rule-based numerical assignment that have little or no epistemic value, such as procedures based on subjective judgment such as (formalized) guesses or statements of opinion, or precisely-specified but arbitrary rules. Views in the second category (focusing on inputs) tend to be too restrictive, insofar as they would disallow many widely accepted cases of measurement in the physical sciences in addition to, potentially, all cases in the social sciences, regardless of epistemic or pragmatic value. Finally, views in the third category (focusing on the formal relationship between inputs and outputs) tend to be incapable of distinguishing processes that generate dependable knowledge from processes that do not but are otherwise functionally similar evaluations.

2.2 Measurement and measurement quality

When considering a set of necessary and sufficient conditions of measurement, intended in its social role as a process able to produce publicly trustworthy information, it may be valuable to note the distinction between two questions: (a) how should measurement be defined, as distinct from other types of evaluations? and (b) how can we judge the dependability of the information provided by an evaluation? In principle, the two questions could be considered redundant if 'measurement' were defined as any evaluation that yielded dependable knowledge, but this would broaden the scope of the concept of measurement to potentially include essentially all forms of inquiry, and is *prima facie* inconsistent with every formal and lay conception of measurement of which we are aware – for example, it is commonly accepted that not all measurements guarantee the same high dependability, and not all opinions are flawed by the same low dependability. Hence, dependability of empirical information is definitely a worthwhile target, but the problem is how such dependability can be obtained and assessed as a stable and publicly justifiable feature of a given process. In other words, measurement per se does not guarantee 'high quality' (whatever this means) information – the concept of 'low quality measurement' is perfectly legitimate. Rather, measurement results are trustworthy because the structure of the process that produces them is such that in principle anyone can assess their quality, however high or low. Stated even more simply, measurement is a source of public trust not because we know *that* we can rely on the information it produces, but because we know *how much* we can rely on it.

The crucial role of quantifier of the quality of the information conveyed by measurement results is played by measurement uncertainty, which is inversely related to information quality: the better the quality the less the uncertainty. This acknowledgment of the structural, and not only operative, importance of measurement uncertainty is relatively new: "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]). 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 a property value but also an estimate of the quality of the information that such a value provides on the property under measurement.

The social prestige of measurement is, of course, the outcome of centuries of development, and thus the novelty of this position is limited. As mentioned, until the recent past measured values were reported together with measurement errors. The relation between measurement error and measurement uncertainty is complicated (for example, some authors simply refer to them interchangeably [Kirkup & Frenkel 2006], thus more or less explicitly denying that there is something new in what has happened in the last decades). From an operative point of view, uncertainty encompasses error: errors generate uncertainty, but uncertainty has sources that are not errors, as in the case of definitional uncertainty. More importantly, the emphasis on uncertainty realizes a conceptual shift, from a purely empirical to a synergetic empirical-informational interpretation of measurement. As a consequence, the central concept of measurement science is arguably no longer the traditionally intended target of a 'true value' that exists independently of measurement and that would be obtained by an error-free empirical process[5] but, indeed, measurement uncertainty (Pane [a] in Figure 2).

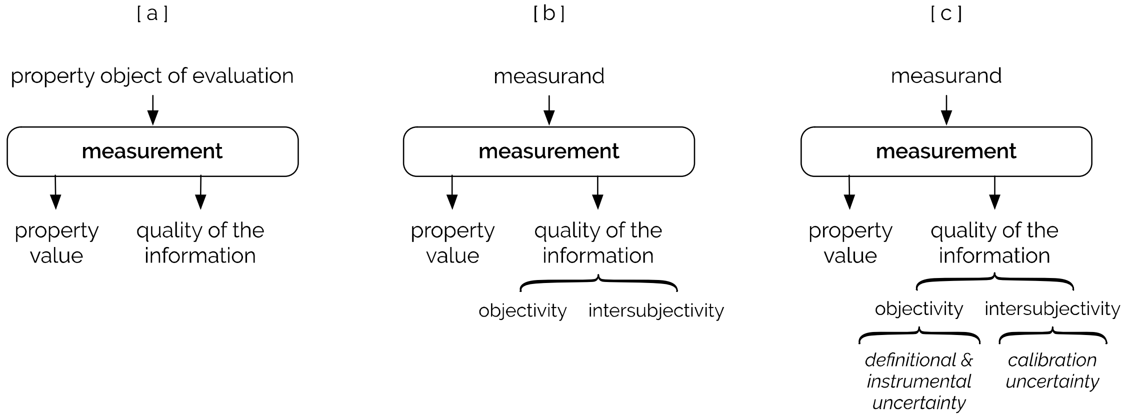

Figure 2. Characterizations of measurement quality.

This paved the way to a fundamental analysis on the nature of the quality of the information conveyed by measurement and on the justification of such quality. In previous works (e.g., [Mari et al 2012]) we have discussed a characterization of the quality of the information conveyed by measurement in terms of two basic features, that we have called object-relatedness and subject-independence, "objectivity" and "intersubjectivity" respectively for short (Pane b] in Figure 2).[6]

Objectivity is intended here as the extent to which the conveyed information is about the property object of measurement and nothing else. The problem of objectivity is twofold. First, empirical properties are interrelated because they are mutually dependent, so that the measurand (i.e., the property intended to be measured [JCGM 2012, def.2.3]) depends on other properties. Since the information produced by measurement is supposed to be usable not only in the moment when it was obtained (and not all the relevant properties might be known in that moment), the issue arises of defining the measurand in a sufficiently specific way so as to make the information transferable without losing the reference to the measurand. Definitional uncertainty is then the means to quantify this component of objectivity. Second, the measuring instrument is generally sensitive not only to the measurand but also to other properties, so

[5] "The input to the measurement system is the true value of the variable; the system output is the measured value of the variable. In an ideal measurement system, the measured value would be equal to the true value." [Bentley 2005].

[6] In philosophy and the social sciences the concepts of objectivity and intersubjectivity have a long history (dating at least back to the 19th century via Edmund Husserl's work on phenomenology [Husserl 1960]), but it would be a mistake to assume that all references to them share the same definition. We use here the terms "objectivity" and "intersubjectivity" in the specific, metrology-related sense introduced elsewhere (see, e.g., [Mari et al 2012]). For an introduction and a discussion of the general philosophical meaning of objectivity, subjectivity, and intersubjectivity, see, e.g., [Davidson 2001].

that its output depends also on them, called “influence 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 so as to make it possible to extract information on the measurand by filtering out the spurious information (“noise”) generated by influence properties. Instrumental uncertainty is then the tool to quantify this component of objectivity.

Intersubjectivity, as intended here, takes into account the goal that the conveyed information is 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 [JCGM 2012, def.2.41] of the information it produces, and therefore the condition of intersubjectivity. Calibration uncertainty, which includes all uncertainties related to the definition of the unit and its realizations in all measurement standards in the traceability chain, is then the tool to quantify intersubjectivity (Pane [c] in Figure 2).

In summary, measurement produces publicly trustworthy information because its objectivity and intersubjectivity are explicitly communicated in terms of measurement uncertainty.

In the context of the traditional paradigm of physical measurement, objectivity and intersubjectivity are features embedded in measuring systems: in other words, measuring systems are designed, set up (including 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 the expected measurement uncertainty This reinforces the point made previously that the public trust afforded to measurement depends not only on knowing *that* the produced information is of high quality, but on knowing *to what extent* this is true. This also further highlights the pragmatic nature of measurement: what counts as high or low quality is relative to the purpose which motivates the measurement; if a comparatively lower quality instrument provides sufficient precision, cheaper measurements may be adopted. Notably, this characterization of measurement quality is independent of any physical condition, and therefore in principle admits realizations also for non-physical quantities (and also for entities that are algebraically weaker than quantities, such as ordinal and even nominal properties: we will not develop this possible extension here). In order to explore how objectivity and intersubjectivity could be assessed in the information obtained in the evaluation of non-physical properties let us then abstract from all physical realizations and focus on the structural features of the measurement process. In reference to the conceptual hierarchy assumed by the VIM, the focus is not on the concrete process (the measurement), nor on the procedure (the detailed description of how the process should be performed), but on the method, the “generic description of a logical organization of operations used in a measurement” [JCGM 2012, def.2.5].

2.3 Three basic methods of measurement

Let us assume that measurement is generically characterized as a process aimed at experimentally obtaining and formally expressing information on a property intended to be measured – the measurand – where the information is reported in explicit relational form relatively to a predefined reference: for continuous quantities the reference is the unit and measurement results are quantity values, hence (sub)multiples[7] of the unit, which actually report the information on the measurand as its ratio to the unit. The fundamental methodological problem of measurement is then how to compare the measurand and the unit.

[7] Strictly speaking, a value such as 1.2 m is neither a multiple nor a submultiple of the metre; a more correct (though cumbersome) expression would be “multiples of submultiples of the unit” (e.g., 1.2 m is the 12th multiple of the submultiple 10^{-1} of the metre).

Three basic methods can be envisaged to this purpose. In order to introduce them in a concrete case, let us suppose that the purpose is to measure the weight W_o (we will not distinguish here between weight and mass) of an object *o*.

Method 1. The object under measurement, of which the measurand is a property, and a measurement standard, which materializes the unit (or standard sequence in the case of ordinal properties), are simultaneously available and a procedure is known to compare them. From the outcome of the comparison a value for the measurand can be obtained. According to this method, W_o could be, for example, compared with each of the weights in a standard series by means of a two-pan balance, and a value for W_o is chosen as the value of the most similar weight of the series. W_o is the only measurand involved here. This method of measurement can be called *direct synchronous*, to emphasize that the measuring instrument directly and at the same time interacts with the object under measurement and a measurement standard.

Method 2. The object under measurement is put in interaction with a transducer that is sensitive to the measurand and produces a new property – the VIM calls it an indication – as outcome. The transducer was calibrated against the unit, by putting it in interaction with a standard that realizes the unit, so that from the indication a value for the measurand can be obtained. According to this method, W_o could be, for example, put in interaction with a calibrated dynamometer (say, a spring balance) and is transduced to a length (the elongation of the spring); a value for W_o is obtained from the length value and the calibration information. W_o is the only measurand involved here. This method of measurement can be called *direct asynchronous*, to emphasize that the measuring instrument directly interacts with the object under measurement and a measurement standard, but this interaction happens in different times.

Method 3. The measurand is related to one or more other properties and the relation is analytically known. Such other properties are measured (according to method 1, or 2, or – recursively – 3) and then the relation is computed on these quantity values to obtain a value for the measurand. According to this method, the volume and the density of *o* could be, for example, measured (according to method 1 or 2, or possibly 3 itself) and from them a value for W_o is computed via the relation weight = volume × density. Three measurands are then involved here: not only W_o, but also two "intermediate measurands", the volume and the density of *o*. This method of measurement can be called *indirect*, to emphasize that there is no empirical interaction with the measurand.

The application of any of these methods of measurement requires that the property intended to be measured – i.e., the measurand – be sufficiently well-defined, where the exact requirements (in terms of clarity, precision, etc) will depend on the purpose for which the measurement is taking place. In addition, each of these methods of measurement is based on both theoretical and operative assumptions: the former are required for the applicability of the method as such; the latter specify the conditions for the quality of the results of measurements performed according to the method. Hence all these methods are theory-laden to some extent (and hence the classical conception of measurement as an atheoretical process, able to produce "pure data", is not compatible with any of them), but with different commitments.

The *direct synchronous method* (Method 1, above) is the least theory-laden, and structurally the simplest. Its only theoretical condition is the experimental comparability between the measurand and the unit, and thus more generally between instances of the same kind of property, a condition that is so basic that it might be even taken as definitional for a kind of property. The most important operative condition of this method is about the measurand selectivity of the comparison, i.e., the requirement that the comparison between the measurand and the unit is not biased by other properties. This explains why the comparison is usually performed by an instrument functionally behaving like the already mentioned two-pan balance.

The *direct asynchronous method* (Method 2) assumes the existence of an empirical transduction effect whose input property is the measurand. In this perspective the theoretical-ladenness of the method may be only related to the hypothesis that the transduction is causal and can be then formalized as a function, mapping the measurand to an instrument indication. Operatively, this method not only assumes that the measurand transduction is not biased by other properties – thus explaining why the transduction is usually performed by an instrument functionally behaving like the already mentioned spring balance – but also that the transducer is properly calibrated and that its behavior is (analytically or numerically) known and invertible.

The *indirect method* (Method 3) requires the application of one or more direct methods on the intermediate measurands and thus assumes their conditions, and additionally assumes that the relation of the measurand with the intermediate measurands is analytically known. The indirection makes this method the most theory-laden, given that a value for the measurand must be computed from values of other, intermediate measurands (in the direct asynchronous method the measurand is related to an instrument indication, which is not a measurand in turn).

The account given up to this point has largely been based on how measurement has developed within metrology – the discipline concerned with the science of measurement and its application – which has, historically, focused primarily on physical quantities, in a strongly structured disciplinary context where the knowledge on general quantities is usually well established, by physics, and underpins the development of measurement.

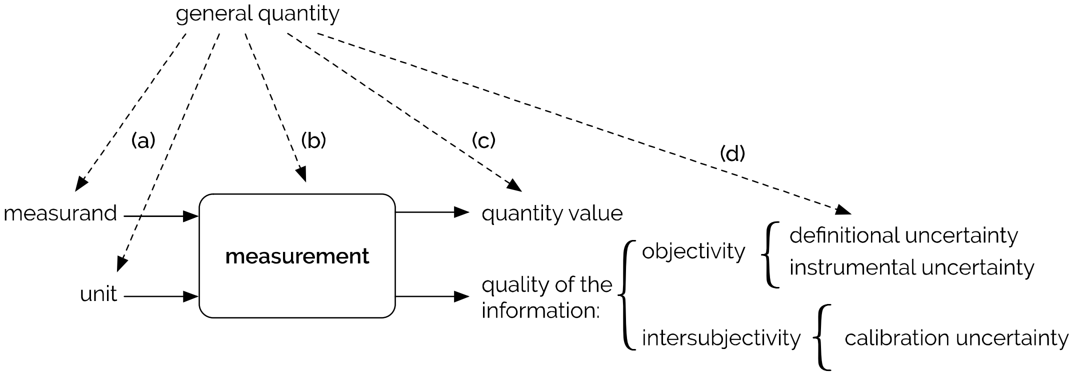

Figure 3. A characterization of measurement in metrology.

In such a context it is assumed that the process is based on previous knowledge on the general quantity:

(a) of which the measurand and the unit are instances,

(b) that the measuring instrument is designed to measure,

(c) of which the quantity value is a possible value, and

(d) whose definition has an uncertainty that is the lower bound of the measurand definitional uncertainty.

This highlights that performing a measurement requires in principle the solution of four preliminary problems:

(1) the general quantity (e.g., weight W) must be defined;

(2) the measurand (e.g., the weight of the object o, W_o) must be defined, as an instance of the general quantity;

(3) the unit (e.g., the kilogram) must be defined, as an instance of the same general quantity;

(4) the measuring instrument must be designed and properly set up so to be able to measure instances of that general quantity.

If the knowledge on the general quantity is sufficiently well established, i.e., there is a good solution to problem 1, as it is usual in physical sciences, the development strategy is top-down: the solutions to problems 2-4 depend on the available knowledge on the general quantity, but not vice versa. Operatively, this is revealed by the fact that for the same general quantity multiple different definitions of measurands and of units are possible and multiple different measuring instruments can be designed (i.e., that the relations (a)-(d) in the figure are all one-to-many): measurements aim at producing information on measurands, whereas the knowledge on the general quantity seldom changes in consequence of this information.

On the other hand, in the preliminary stages of the definition of a general quantity it may happen that the definition of its individual instances is not clearly distinguished from the definition of the general quantity itself (problems 2 and 3), which is based on the design of a specific measuring instrument (problem 4), so that the development strategy is unavoidably bottom-up and the knowledge on the general quantity is improved by measuring some of its instances.

In the following section we attempt to apply the analysis introduced here to the social sciences – where measurement models and methods are usually less established than in physical sciences – with two main purposes: first, as a way of checking whether and under which conditions our fundamental claim, that measurement produces publicly trustworthy information because its results contain explicit information on their own quality, applies also to the processes that are considered to be measurements in social sciences; second, as a way of evaluating the coherence and justifiability of measurement practices in the social sciences.

3. The meta-structural understanding of measurement in the social sciences: a discussion

Over approximately the last century measurement has become an integral concept in many areas of the social sciences. For example, it is regularly claimed or implied that scores on academic tests or surveys constitute results of measurements of knowledge, skills, and abilities, attitudes, motivations, perceptions, and personality factors. It is rare to encounter a social scientist who does not believe not only that the measurement of human attributes is possible in principle, but also that it has been achieved in practice. Given the importance of the social consequences of test use throughout the social sciences (e.g., [Hornstein 1988], [Porter 1996]), in addition to the importance of measurement quality for scientific inquiry in general, justification of the quality of measurement results would seem to be at least as important a topic in the social sciences as it is in other settings.

As we and others have argued elsewhere (e.g., [Mari et al 2016], [Cano et al 2016]), in principle, there does not seem to be any a priori reason why the justification of the quality of measurement results should take substantially different forms for physical and non-physical properties. According to our understanding, what is sought, in both cases, is justification of claims of objectivity and intersubjectivity, or in other words that measurement results are related only to the property of the object under consideration, and not other influences, and that information obtained through measurement is interpretable in the same way by different persons in different places and times. Importantly, the objectivity and intersubjectivity of measurement results do not require that the measured property be definable in purely physical terms or exist independently of human consciousness; in other words, there is no contradiction in the claim that a measurement procedure can yield epistemically objective information about an ontologically subjective phenomenon [Maul 2013]. In the social sciences, the goal of objectivity has been described variously as a need for a lack of "construct underrepresentation" and "construct-irrelevant variance" (e.g., [Messick 1995]), a need for "measurement invariance" to factors not related to the measurand (e.g., [Rasch 1960]), and a need for there to be a causal relationship between variation in a measurand and variation in the outcomes of a measurement process (e.g., [Borsboom et al 2003]). By contrast, the topic of intersubjectivity, as understood here, has received less explicit attention thus far in the social sciences, plausibly also due to the fact that the social sciences lack something analogous to a metrological system, made of quantity units, measurement standards, and traceability chains though this is not to say that these issues have not each received attention from scholars in the social sciences, of course (e.g., [Humphry 2011]).

Thus, in principle, our argument that measurement is an experimental process characterized by its structure and able to provide explicit information on the quality of its own results should apply to the social sciences as well. With this in mind, it may be informative to examine how the distinctions between methods of measurement introduced in the previous section, and more generally our claim that measurement is an evaluation that provides both property values and an assessment of their quality, apply to measurement in the social sciences.

3.1 The three basic methods of measurement in the social sciences: an example

For illustrative purposes, we suppose that a team of researchers is interested in measuring the political orientation (henceforth, PO) of individuals. We assume that these researchers seek to gather information on PO by means of a questionnaire for basic research purposes only, and thus with no intended consequences at the level of the individual. As a general property, 'political orientation' is defined as the overall extent to which an individual's beliefs are consistent with a conservative versus a progressive ideology, where 'conservative' and 'progressive' are defined with reference to the literature on political theory. PO may be characterized as either an ordinal or a quantitative property of persons, whose poles are then progressive and conservative. The assumption that PO is only ordinal has the consequence that the reference to which the measurand has to be compared is not a unit but a sequence of individual POs. An instrument (to be described in more detail in each of the following examples) is built for the purpose of experimentally obtaining and formally expressing information on the political orientation of individuals. As for the physical example above, the whole process can be characterized in black box terms, by defining political orientation as the input of a transformation process that eventually produces a property value.

political orientation
transformation
property value

Figure 4. Black box measurement of political orientation.

As was argued previously, the trust afforded to measurement results depends on the structure of the process that produces them. How would the quality of the results of the measurement of PO be justified then? To answer such a question, we need first to reconstruct the structure of the process in the terms given in the previous section. Our attempt at reconstruction contains three sub-questions:

(a) how can a reference PO be defined? this is about an individual property, in analogy with identifying a measurement unit in the case of a quantity;

(b) how can a standard of PO be identified or realized? this is about an object having a reference property, in analogy with identifying a measurement standard in the case of a quantity;

(c) how can a person and a standard of PO be compared to one another in terms of PO?

In what follows, we will explore the tenability of understanding the structure of the process of measuring PO in terms of each of the three basic methods of measurement described previously. In each case, we comment on the extent to which the situation described coheres with established concepts and practices in the social sciences, as well as the defensibility of the proposed interpretation.

Direct synchronous measurement of PO. Since PO is a property of human beings, in principle, it might be possible to identify some number of "prototypical" individuals and to define their POs as the reference POs: for example, three individuals could be defined as instantiating progressive, neutral, and conservative POs respectively. The task of an instrument would then be to support the comparison of individuals subject to measurement with the prototypical individuals with respect to their PO, and find the one with whom the PO of the individual is most similar (discussion of uncertainty omitted here for simplicity). This solution has the same structure as the Mohs scale of hardness, with human beings in place of minerals and PO of human beings in place of hardness of minerals.

Of course, such a process would require specification of the method for comparison of individuals, to be then implemented in the instrument. In contrast to physical properties such as hardness and weight, PO is not transparently the sort of property that admits of direct synchronous comparison.[8] Any instrument designed for the purpose of comparing individuals' POs would instead require some specification of how individuals' POs are related to (some set of) observations, thus introducing asynchronicity into the process.

Direct asynchronous measurement of PO. A questionnaire could be devised with a series of short, declarative statements expressing various strengths and severities of progressive and conservative beliefs, asking responding individuals to indicate whether they would endorse each statement, or the extent to which they would endorse each statement given a set of response options. The questionnaire operates in this case as a transducer that interacts with the PO of the human being who responds and produces a response pattern as an outcome. Instead of directly taking it as the indication, this response pattern may be then mapped into another value, such as the raw score (e.g., for social scientists working from a classical test theory perspective), or an estimate of a location along a 'latent variable' (e.g., for social scientists working from a factor-analytic or item response theory perspective). The function that takes response patterns and associates them with, e.g., raw scores may be generically called a "synthesis map", its output being then interpreted as the indication value.

The problem of how to calibrate the questionnaire as a transducer is handled as an iterative process, iterating between working on the definition of the general quantity, the set of questions that are used in the transducer, and checking that the questions are operating in a manner consistent with the intentions of the test designers. (This stands in contrast to physical measurement instrument calibration, which is usually unrelated to the definition of the general quantity, and instead assumes it.) Such a process generally proceeds by a method of successive approximations. There is more than one tradition of this (see [Wilson 2013] for an account of several). For example, one such method, called *construct mapping*, proceeds by assuming that there is a sufficiently clear definition of the general quantity that one can order the questions (to a given degree of certainty) with respect to their relationship to the general quantity itself, so that the order is empirically testable by comparing it with the order of the estimates of item parameters. The method is made theoretically more stringent, and simpler, by specifying that the resulting questionnaire should also demonstrate "specific objectivity" [Rasch 1960]—that is that the results for a measurand should be equivalent no matter which items had been used from within a certain set (analogous to the requirement that the results of measuring mass using a two-pan balance should be equivalent no matter which balance had been used), thus setting up the possibility of a definition to a universe of items. This approach also leads to a possibility of collecting evidence relevant to both instrumental uncertainty (via standard errors of estimates of person locations), and definitional uncertainty (via person misfit statistics).

As for all measurements based on a direct asynchronous method, the indication needs to be then related to a value of PO, and this requires calibrating the questionnaire, i.e., constructing its transduction function by obtaining the (functions of) response patterns generated by applying the questionnaire to the elements of a sequence of standards, and then analytically or numerically inverting such a function to obtain a calibration function or map. In principle, this calibration could take place using a number of different methods, but in practice, in the social sciences, this step is usually not given explicit attention.

Here, PO is defined as a property of persons (or a "construct", or a "trait", among other possibilities) that is causally responsible for behaviors (including expressions of attitudes) across a wide variety of situations, both experimentally observed and not. If the questionnaire functions as intended, (between-person variation in) PO causes (between-person variation in) responses to questionnaire items [Borsboom et al 2003]; thus, the items act as a specific transducers between PO, considered an unobservable construct, and observable responses to items. The transduction is acknowledged to be fallible, as variation in individuals' response patterns may be subject to additional influences (which could be termed "influence properties" whose presence causes "construct-irrelevant variance"), such as the language abilities or level of alertness of the person responding to the items. A variety of approaches have been developed within latent variable frameworks for quantifying the nature and extent of the impact of multiple influences on

[8] Since the first stage of knowledge of physical properties typically is the direct synchronous comparison of their instances (this is longer than that, this is warmer than that, etc), the fact that the development of measurability of non-physical properties usually skips this step and instead starts from their tentative mapping to numbers / symbols, thus "jumping to the second rung of the ladder", might be considered a reason for the sometimes weaker empirical ground of measurement of non-physical properties. This seems to be a subject worthy of exploration in future works.

response patterns (e.g., [de Boeck & Wilson 2004], which might be interpreted as methods for evaluating instrumental uncertainty.

In the social sciences, directed acyclic graphs (DAGs) [Tu 2012] are often used to represent latent variable models. Such diagrams are composed of three kinds of objects: circles, rectangles, and arrows. The circles are used to represent unobserved properties ('latent variables'), the rectangles represent properties that are directly observed (such as item response data), and arrows represent hypothesized causal relationships (with the nature of these relationships left unspecified, or in other words, as black boxes). There are many distinct types of DAGs, two of the most common and relevant of which are briefly described [Edwards & Bagozzi 2000]. In a 'reflective' model (Pane [a] in Figure 5), the unobserved property (in the present case, PO) is hypothesized to be causally responsible for observable behaviors (in the present case, responses to questionnaire items). In a 'formative' model (Pane [b] in Figure 5), the unobserved property is instead modeled as an inductive summary of the observations, caused by them, in the (relatively trivial) sense in which any summary is caused by the facts being summarized. Despite the similarity in diagrams and the corresponding models, the situations depicted are importantly distinct: while the reflective model here represents a situation in which political orientation is measured via responses to questionnaire items, the formative model represents a situation in which political orientation is simply defined as a summary – or more specifically, as a weighted combination – of the questionnaire items. As such, given the perspective on measurement described previously in this paper, reflective models are regarded as *measurement* models, whereas formative models are devices for the reduction and summarization of data [Rhemtulla et al 2015]. More generally, as has been previously argued, measurement is a process aimed at experimentally obtaining and formally expressing information on a property intended to be measured, whereas summarization as such does not aim at obtaining new information; put even more simply, measurement is inferential, whereas summarization is only descriptive.[9]

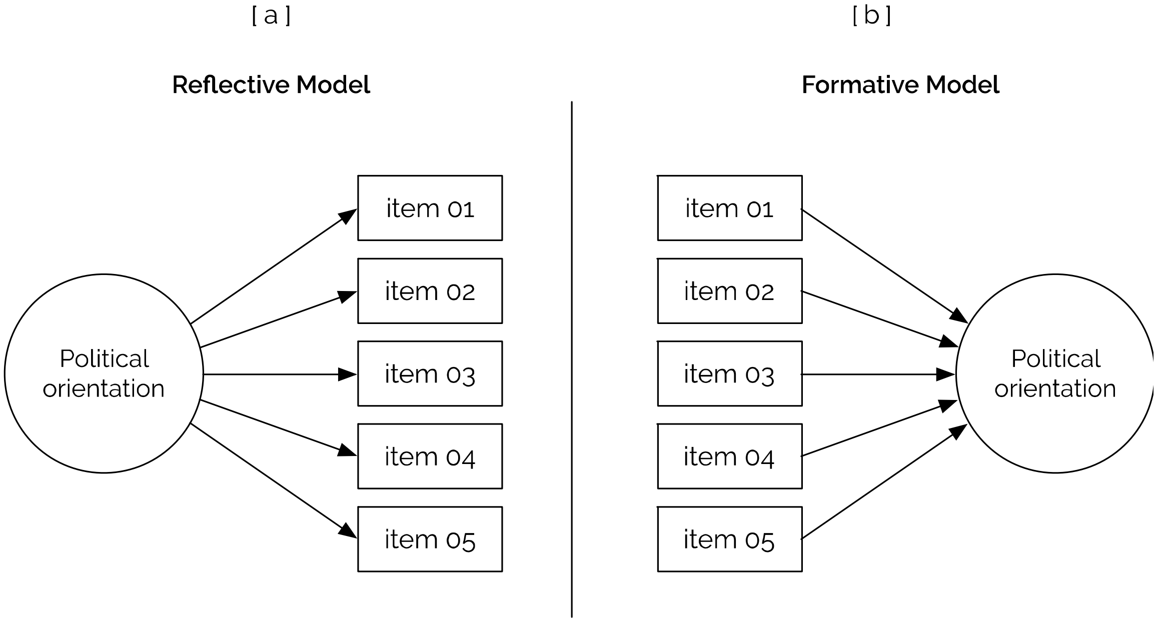

Figure 5. Directed acyclic graphs for reflective and formative latent variable models.

In the present context, maintaining the formative perspective that PO is *nothing more than* a (weighted) combination of responses to survey items does not seem plausible, for the simple reason that one's political orientation is generally thought to be a characteristic of individuals with implications for their beliefs and

[9] The points made here also reinforce the position that not every procedure that yields a number is a measurement. In the social sciences, many procedures for generating numbers are data-reduction devices rather than measurements.

behaviors in many environments (and thus not only responses to questionnaire items). A more typical example of a formative model would be one in which an individual's socio-economic status (SES) is defined as a combination of specific facts such as their income, net worth, education level, and geographic location: in this case, SES is not thought of as something that exists separately from and causes variation in these facts, but rather, as a summary of them.

Still, this characterization of the process of measuring PO is not fully consistent with the way many scholars and practitioners would describe the situation: in particular, there are many who would instead emphasize that terms like "political orientation" are used to summarize and describe a larger set of more specific facts about an individual (e.g., in terms of beliefs, attitudes, behavioral dispositions, etc.) – facts that stretch beyond the immediately-observed responses to questionnaire items. This perspective could be considered more consistent with the position that the questionnaire should be considered a tool for indirect measurement.

Indirect measurement of PO. A person's PO could be defined as a combination (or "composite") of several more specific beliefs and attitudes, such as attitude about taxation (*X*), attitude about abortion (*Y*), and attitude about privatization of education (*Z*), presumably among a long list of others, thus assuming that PO is not exhaustively defined by *X*, *Y*, and *Z*. We then might suppose that this definition helped motivate the writing of questionnaire items, with the intent of measuring not PO as such but each of these more specific beliefs and attitudes, thus intended here as intermediate measurands. Hence the measurement of PO first involves the measurement of these more specific attitudes, which could take place for example using the direct asynchronous methods as described previously. In a second step, knowledge of how these specific attitudes are sub-components of PO, as a result of the definition of PO, is leveraged to compute a value of PO – which might take place, for example, via a hierarchical latent variable model, involving reflective models for the measurement of *X*, *Y*, and *Z*, and a formative model for their combination, as visually represented by the DAG in Figure 6. In this scenario, the intermediate measurands *X*, *Y*, and *Z* have directly interacted with the measurement instrument, and by means of computation a value of PO has been estimated; thus this could be considered a case of indirect measurement.

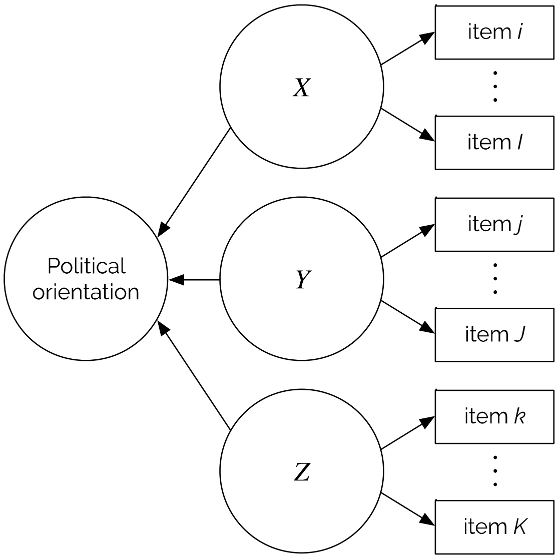

Figure 6. A hierarchical latent variable model.

These diagrams could be rearranged to be more consistent with those presented earlier, as in Figure 7. Uncertainties in the measurement of *X*, *Y*, and *Z* are evaluated as before and then propagated through the composition map, which, given that here PO is defined as a combination of *X*, *Y*, and *Z*, might add a definitional uncertainty.

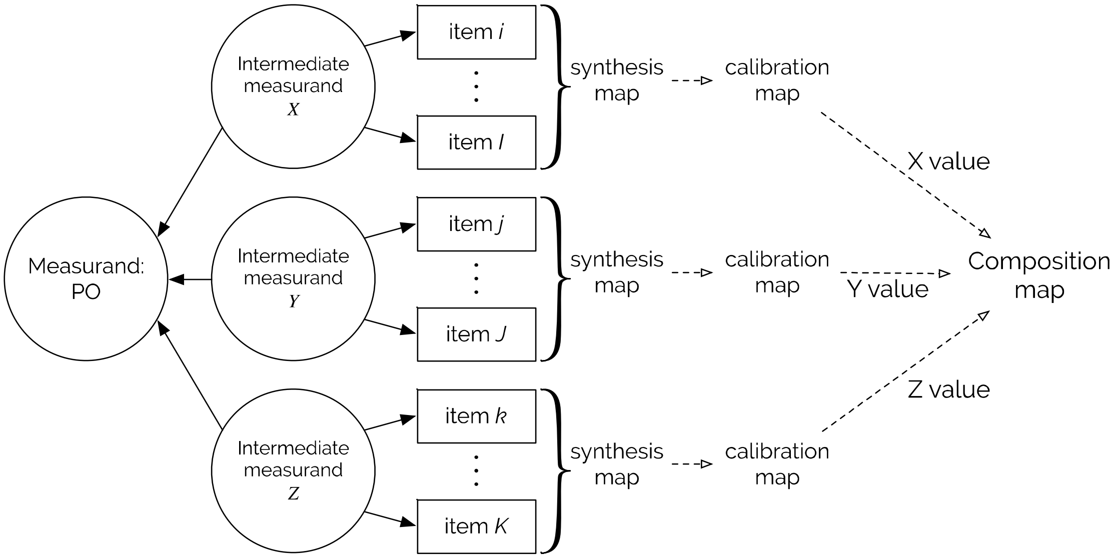

Figure 7. Indirect measurement of political orientation.

It could be further noted that this sort of logic seems to be present in a wide range of applications in the social sciences (for one exposition of this position, see, e.g., [Snow & Lohman 1989], whether or not the two-stage process described here (i.e., direct asynchronous measurement of sub-components followed by an analytic combination of these sub-components) is formally followed. In many instances that we see in the literature, there may be only a single item (or very few items) associated with each sub-component of PO, and thus the estimation of PO will be collapsed and will not involve two explicitly separate steps. Instead, the estimation might take place using a standard latent variable model, despite the relationship between the property and the sub-components of the property measured by the individual questionnaire items being thought to be one of constitution rather than causality.

3.2 The definitions of general properties in the social sciences

Throughout the examples given in the previous section, it can immediately be noticed that ambiguity in the definition of 'political orientation' itself has obfuscated the efforts to reconstruct the structure of the process of its measurement. This problem – ambiguity in the definition of the general property to be measured – may not be unique to the social sciences, but clearly plays a much more prominent role in measurement in the social sciences than it does in physical metrology: as previously discussed, in the physical sciences general properties (e.g., length, mass, energy) are usually already defined prior to the construction of a new measuring instrument. In contrast, in the social sciences it is common for the definition of a general property, or "construct", to be largely formulated interactively within the act of instrument construction. In more sophisticated cases, the specification and elaboration of a definition of a property may take place iteratively over many cycles of instrument development and refinement, in dialogue with substantively-oriented researchers (e.g., political psychologists, in the previous example), and supported through triangulation of several different candidate measurements of a single property. Often, however, constructs are given ad hoc, ostensive definitions in the context of a given instrument, with the frequent consequence that the same term is used to refer to distinct properties (or, even more problematically, that it is not clear to what extent this is the case).

If it were possible to define PO in terms of "prototypical" individuals, it could in principle be possible to reconstruct a relevant measurement process as an instance of direct synchronous measurement, though, as previously argued, this is implausible given the non-physical nature of the property. If PO were instead defined as a 'latent' property of persons that causes item responses, the process could be reconstructed as one of direct asynchronous measurement, as the questionnaire items and the latent variable model act respectively as transducers and synthesis maps between political orientation and the outcomes of the measurement process. Finally, if PO is instead defined as a composite of several more specific properties of a person, a further computational layer is added, in which case the process could be reconstructed as one of indirect measurement, based on the social science version of direct asynchronous measurement.

The fictional example used here comports with our experiences of common practices in the social sciences: specifically, it is common for general properties such as PO to be given definitions too ambiguous for it to be possible to unequivocally reconstruct the structure of the measurement process in anything more than black box terms. This, in turn, severely hampers our efforts to evaluate how structural features of the measurement process could guarantee the quality (i.e., objectivity and intersubjectivity) of the results.

4. Discussion

In this paper we have argued that justification of the dependability of measurement results depends on the structure of the measurement process, rather than (solely) on features of the inputs or outputs of the process (such as whether or not they are quantitative), or the functional relationship between the two. As such, efforts to understand the extent to which any given application of measurement produces results of sufficiently high quality for a specified purpose requires "opening the black box" of the process, and understanding the manner in which features of the process relate input properties to outcomes. We have argued that measurement quality can be understood in terms of objectivity and intersubjectivity, and discussed how operative features of the measurement process are related to the quantification of measurement quality in terms of measurement uncertainty (including definitional uncertainty, instrument uncertainty, and calibration uncertainty).

Then, as a stepping stone towards considering how measurement quality could be assessed in the evaluation of non-physical properties, we discussed how structural features of the measurement process could be abstracted from their physical realizations, and presented three basic methods of measurement in structural terms.

Following this, by means of an example we explored how the measurement of a non-physical property could be reconstructed in these terms. In doing so, we noted three important differences between instances of measurement typical of the social sciences and those typical of classical metrology. First, in the social sciences, measurement standards (i.e., units in the case of quantities, or reference sets in the case of ordinal properties, disseminated through traceability chains) are, for the most part, absent (or, where present, are idiosyncratic to particular contexts, thus providing little basis for universally accessible communication), so that *the metrological traceability of data claimed to be measurement results might be hard to justify*. Second, the definitions of non-physical properties are typically not as well formulated as the definitions of most physical properties, thus *leading to big, even though usually implicit,* definitional uncertainty, and also hampering efforts to provide clear reconstructions of the structures of measurement processes in the social sciences. Third, and partly as a consequence of the second point, efforts to develop instruments in the social sciences are often inextricable from efforts to define general properties (and obtain information to help guide the formulation of such definitions) – and indeed, *definitions of general properties commonly change over the course of a given effort to develop a new measuring instrument* in the social sciences – further challenging efforts to reconstruct non-physical measurement processes. However, all three of these differences could be interpreted as historical and contextual rather than essential: that is, none of these three differences suggests that the process of measurement is fundamentally different for physical and non-physical properties.

References

Allen, M.J., & Yen, W.M. (1979/2002). *Introduction to measurement theory*. Prospect Heights, Illinois: Waveland Press.

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

Berglund, B., Rossi, G.B., Townsend, J.T., & Pendrill, L.R. (editors) (2011). *Measurement with persons: Theory, methods, and implementation areas*. Psychology Press.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review, 110*, 203.

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

Boumans, M., Hon, G., & Petersen, A.C. (editors) (2013). *Error and uncertainty in scientific practice*. Routledge.

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

Cano, S.J., Vosk, T., Pendrill, L.R., & Stenner, A.J. (2016). On trial: the compatibility of measurement in the physical and social sciences. *Journal of Physics: Conference Series, 772*, 1.

Davidson D. (2001). *Subjective, intersubjective, objective: Philosophical essays Volume 3*. Oxford: Clarendon.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

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

Edwards, J.R., & Bagozzi, R.P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Hornstein, G. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J.G. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 1-34). New Haven: Yale University Press.

Humphry, S.M. (2011). The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 9*, 1-24.

JCGM 100:2008, *Evaluation of measurement data – Guide to the expression of uncertainty in measurement* (GUM, originally published in 1993), Joint Committee for Guides in Metrology, 2008 (http://www.bipm.org/en/publications/guides/gum.html).

JCGM 200:2012, *International Vocabulary of Metrology – Basic and general concepts and associated terms* (VIM), 3rd Edition (2008 version with minor corrections), Joint Committee for Guides in Metrology, 2012 (http://www.bipm.org/en/publications/guides/vim.html).

Kirkup, L., & Frenkel, R. (2006). *An introduction to uncertainty in measurement*. Cambridge: Cambridge University Press.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1971/1989/1990). *Foundations of measurement, vols. 1–3*. New York: Academic Press.

Kyburg Jr., H.E (1984). *Theory and measurement*. Cambridge: Cambridge University Press.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. London: Addison-Wesley.

Margenau, H. (1958). Philosophical problems concerning the meaning of measurement in physics. *Philosophy of Science, 25*, 23–33.

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 Trans. Instr. Meas., 61*, 2107–2115.

Mari, L., Maul, A., Torres Irribarra, D., & Wilson, M. (2016). A meta-structural understanding of measurement. *J. of Physics: Conf. Series,* 772.

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

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79,* 311-320.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, N.J: L. Erlbaum Associates.

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

Narens, L. (1985). *Abstract measurement theory*. Cambridge: MIT Press.

Narens, L. (2013). *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science*. Psychology Press.

Narens, L. (2014). *Theories of meaningfulness*. Psychology Press.

Pfanzagl, J. (1968). *Theory of measurement*. Wurzburg: Physica-Verlag.

Porter, T.M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, N.J: Princeton University Press.

Porter, T.M. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinary Research & Perspective*, *1*(4), 241–255.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, Ill: Univ. of Chicago Press.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2015). Calling models with causal indicators measurement models implies more than they can deliver. *Measurement, 13,* 59-62.

Roberts, F.S. (1979). Measurement theory – With applications to decision-making, utility and social sciences. In: G. Rota (Ed.), *Encyclopedia of mathematics and its applications, vol. 7*. London: Addison-Wesley.

Sawyer, K., Sankey, H., & Lombardo, R. (2016). Over-measurement. *Measurement*, *93*, 379-384.

Schlaudt, O., & Huber, L. (editors) (2015). *Standardization in measurement: Philosophical, historical and sociological issues*. Routledge.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic, 23*, 2, 113–128.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational testing. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 263–331) New York: American Council for Education.

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

Tu, Y.K. (2012). Directed Acyclic Graphs and Structural Equation Modelling. In *Modern Methods for Epidemiology* (pp. 191–203). Springer Netherlands.

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