preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

Intersubjectivity of Measurement Across the Sciences

Andrew Maul[1#], Luca Mari[2], Mark Wilson[3]

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

Abstract. A critical condition for the quality of measurement results is that they be interpretable in the same way by everyone, even though they may have been obtained in different contexts by different individuals using different instruments: in other words, they should be subject-independent, or intersubjective. For both physical properties and psychosocial properties, intersubjectivity can be secured by establishing the metrological traceability of measurement results to a measurement unit, and more generally to a set of reference properties, though at present such solutions are less commonly found in psychosocial applications. In this paper we describe traditional and newer solutions to the problem of intersubjectivity in the physical sciences, and then explore how these and other solutions can apply to non-physical measurement as well. The fact that, despite their differences, the metrological traceability to references can be structurally guaranteed in both physical and non-physical measurement and can be presented in a single and consistent framework is a significant step towards the development of a conception of measurement across the sciences.

Keywords. measurement; philosophy of measurement; intersubjectivity; units, measurement standards; reference properties

1. Introduction

We begin by asking the reader to imagine the following scenario. Suppose that two individuals, in two different places and times, using two different instruments, each measure the hardness (considered as an ordinal property) of two different objects. That the first individual finds the hardness of her object to be “8”, and the second individual discovers the hardness of his object to be “7”; both individuals then report their results, referring to the same scale. Under what conditions would we be able to confidently conclude that the hardness of the first object is actually greater than the hardness of the second object?

We might also imagine a second scenario, identical to the first in structure, except that instead of measuring the hardness of two physical objects, we were instead concerned with evaluating the chess playing ability of two individuals. Again we could ask the question: under what conditions would we be able to confidently conclude that the chess playing ability of the first person is actually greater than the chess playing ability of the second person?

Further, does anything about the answers to these questions change if we consider instead properties commonly modeled as quantitative rather than ordinal, such as length, temperature, or reading comprehension ability (e.g., Pearson & Hamm, 2005)?

The scenarios described here highlight a critical issue regarding the quality of measurement results, and both pose a special case of a more general question: under what conditions are we willing to accept that measurement results are of high enough quality to be depended upon? As we have argued elsewhere (Maul, Mari, Torres Irribarra, & Wilson, 2018), the societal role of measurement is underwritten by trust in the quality of the information obtained. In particular, we proposed that there are two major dimensions of measurement quality

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

in need of credible documentation: object-relatedness, or *objectivity*, and subject-independence, or *intersubjectivity*. From our perspective, then, the primary task facing anyone wishing to justify the dependability of measurement results—regardless of whether the measurement is of a physical property such as hardness or a non-physical property such as chess playing ability—is documentation of the structural features of the measurement process that serve to secure both objectivity and intersubjectivity.

To perhaps state the obvious, this is no trivial task. The framing of the two scenarios just presented was intended to call attention in particular to one aspect of the overall task, related to the credible documentation of the features of the measurement system that allow measurement results obtained by different individuals in different times and locations and using different instruments to be meaningfully compared—in other words, to be subject-independent, or intersubjective. In order to focus on the requirement of intersubjectivity and the ways in which it may be secured, we assume here that objectivity has already been secured—in other words, we assume that the information obtained on hardness and chess playing ability (and other examples used in this paper) truly does relate to that property of the object under measurement, which in turn requires the assumptions that (a) the properties in question have been sufficiently well-defined, making definitional uncertainty (JCGM, 2012: 2.27) negligible, and (b) that the measurement instruments were not sensitive to other properties (also termed "influence properties"), making instrumental uncertainty negligible. These are clearly not trivial assumptions. For now, though, committing to them allows us to focus on the following two primary questions:

(1) How can intersubjectivity be structurally guaranteed?, and
(2) How does the answer to the first question change, depending on the area of application (e.g., physical properties, psychosocial properties, etc.)?

This paper is structured as follows. In Section 2 we first say more about the problem itself, to attempt to clarify what is at stake. We then briefly review in Section 3 solutions found in physical metrology to guarantee intersubjectivity, followed in Section 4 by a similar review of solutions that have been or could be proposed in the human sciences. We conclude with a discussion of the similarities and differences of these solutions. We note some issues that condition the possibility of guaranteeing subjectivity in the human sciences, but argue that, despite their differences, traceability can in principle be structurally guaranteed in both physical and non-physical measurement settings and can be presented in a single and consistent framework, which we propose is a significant step towards the development of a single, consistent conception of measurement across the sciences.[1]

2. The problem to be solved, the conditions to be guaranteed

Measurement results convey information on the relation between the measurand (JCGM, 2012: 2.3) and one or more reference properties. In the canonical case of additive quantities, this

[1] The more general issue of the comparability of measurement concepts and practices across different areas of application has been the subject of a significant amount of scholarship over the past century. This is 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), Schlaudt & Huber (2015).

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

relation is between the measurand and another quantity of the same kind, chosen as the quantity unit. By stating that, for example, the length $L[a]$ of rod a is 1.23 m, possibly together with some measurement uncertainty, we claim that the two lengths identified as $L[a]$ and the metre have been compared and the former is 1.23 times greater than the latter, i.e., that $L[a] / \mathrm{m} = 1.23$. Property evaluations reported in reference to interval, ordinal, and even nominal scales[2] are analogous: their results are relational, again involving a measurand and other reference properties; only the structure of the involved relation changes. For example, the result that the temperature $T[a]$ of rod a is 23.4 °C means that both a unit and a zero temperature have been conventionally chosen – let us designate them as u_C and z_C respectively – and $(T[a] - z_C) / (u_C - z_C) = 23.4$. Likewise, the result that the hardness $H[a]$ of a is 7 in the Mohs scale conveys the information that the hardness of a is equivalent to that of the seventh element of the scale, i.e., quartz.

In all of these cases, before the measurement a set of reference properties or quantities is established (and the scale type determines how this can be done), and the measurement then consists of the (direct or indirect, explicit or implicit) comparison of the measurand with the elements of the set. The fact that measurement results are in this sense relational has the consequence that properties of objects become comparable not only by empirically comparing objects by their properties but also mathematically via the values of these properties. For example, if $L[a] = 1.23$ m and $L[b] = 2.34$ m, then the ratio $L[b]/L[a]$ is scale invariant, and therefore we can infer that $L[b] = 2.34/1.23\ L[a]$, even without directly comparing a and b by their lengths. Analogous conclusions can be drawn in the interval and ordinal cases mentioned above: if $T[a] = 23.4$ °C and $T[b] = 34.5$ °C, then the ratio $(T[b]–z_C)/(T[a]–z_C)$ is scale invariant, and therefore we can infer that $T[b] = 34.5/23.4\ (T[a]–z_C) + z_C$, even without directly comparing a and b by their temperatures; and if $H[a] = 7$ Mohs and $H[b] = 8$ Mohs, then the relation $H[b] > H[a]$ is scale invariant, and therefore we can infer that $H[b] > H[a]$, even without directly comparing a and b by their hardnesses.

The validity of these inferences depends on a number of premises. First, as stated previously, we assume that the measurement results are objective, in the sense of referring to the property in question (hardness, temperature, length) and not to anything else. This also involves the condition that it is in fact possible to consistently measure the considered properties on the assumed scale types (and therefore that lengths can be measured on a ratio scale, and so on), a nontrivial premise that we will not comment further on here (though see Mari, Maul, Torres Irribarra, & Wilson, 2017). A second premise, more hidden but not less critical (and in fact the one we mainly explore here), is that the scale against which the two measurands were compared (i.e., the scale generated by the metre in the ratio example) is the same. It should be noted that this is unrelated to the linguistic choice of using the same term ("metre" or whatever else) for referring to the scale in the two measurement results: since quantity units are quantities in turn, and more generally reference properties are properties in turn, this is an empirical hypothesis and as such it must be justified. The actual inference therefore has this structure[3]:

[2] By "scale" (or "reference scale"), we mean an appropriate set of individual properties that are identified as properties of objects, not a set of values of a property. (Values of a property result from the scale, not vice versa: trivially, we first define the metre then we assign the value 1 m to it). Also, following the tradition of Stevens (1946), we use the terms "scale" and "scale type" even for nominal properties, for which the concept 'scale', which recalls ordering, would not be applicable. We accept here that measurability is not a priori constrained by scale type. A justification of this position is given in Mari, Maul, Torres Irribarra & Wilson (2017).

[3] Interestingly, the structure of this inference is the inverse of the structure of measurement according to the representational theory of measurement (e.g., Krantz, Suppes, & Luce, 1971, 1980, 1990): here the empirical

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

Premise 0: A given *property* can be measured on a given *scale type*
Premise 1: the *property* of object *a* has been measured to be v_1 with respect to the scale s_1
Premise 2: the *property* of object *b* has been measured to be v_2 with respect to the scale s_2
Premise 3: $s_1 = s_2$
Conclusion: the formal relation[4] between v_1 and v_2 corresponds to the empirical relation between the properties of *a* and *b*

The three cases mentioned above immediately fit into this template. For example:

Premise 0: length can be measured on a ratio scale
Premise 1: $L[a]$ has been measured to be 1.23 times a given length unit u_1
Premise 2: $L[b]$ has been measured to be 2.34 times a given length unit u_2
Premise 3: $u_1 = u_2$
Conclusion: $L[b] = 2.34/1.23\ L[a]$

And analogously for temperature and hardness. In all cases, Premise 3 states the condition that the reference to which $P[a]$ was compared is the same (or could through an appropriate function be made the same[5]) as the reference to which $P[b]$ was compared, even though the two comparisons might have been performed in different places and times. Again, this is an empirical hypothesis and as such it must be justified. Providing such a justification is then an important task, as it establishes the *metrological traceability* of measurement results, possibly produced in different contexts, to the same reference scale (and therefore to a unit in ratio cases, or to a unit and a zero in interval cases, or to a reference sequence in ordinal cases): in particular, the same result should refer to the same empirical situation independently of where and when and by whom it was obtained. This guarantees the metrological comparability of measurement results (JCGM, 2012: 2.46), that makes them identically interpretable by different measurers: that is, intersubjective. Thus, intersubjectivity is one of the characterizing features of measurement, and helps explain its societal role.

The intersubjectivity of measurement depends on the traceability of its results to the same reference scale, as in Premise 3 in each of the arguments above, which in turn depends on the availability of the scale in the context where measurement is performed. In principle, this should identically apply to both physical and psychosocial properties. In a globalized society, in which measurements may be performed at any time and location, this can be a challenging condition to guarantee. The claim that there are recognizable, fundamental features that are common to (what is commonly claimed to be) physical and psychosocial measurement is premised on the solution of this problem. Hence, we first briefly describe the evolutionary process followed in physical measurement to guarantee intersubjectivity, and then we compare it with some solutions proposed regarding measurement of non-physical properties.

relation of properties is the result of sufficiently intersubjective measurements; in RTM it is a condition to construct measurements.

[4] It could be noted that there is another trivial premise at work in these examples, which is that the numbers appearing in the measurement results have formal properties of their own: e.g., in the ordinal cases, that $8 > 7$; in the ratio and interval cases, that two numbers have a given ratio, etc. In general such a premise need not be stated, but making it explicit helps make clear the idea that the conclusion is the (non-trivial) empirical analogue of the (trivial) formal relation between the numbers involved.

[5] A more general version of Premise 3 is indeed $s_1 = f(s_2)$, where f is a known function (e.g., to convert from metres to inches).

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

3. Solutions found in physical metrology

3.1. The traditional solution in physical measurement

In the context of physical measurement, the traditional solution has been based on the strategy of identifying a reference property as the property of a given object. In the case of quantities that are invariant on a ratio scale, the unit is then defined[6] as the individual quantity of a given object, which is assumed to be stable and is referred to as the *primary measurement standard*. In the example of length:

$$\mathrm{m} := L[s_0]$$

where m is the unit (e.g., the metre) and s_0 is the primary standard, which might, for example, be a given (suitably produced and maintained) rod, whose length $L[s_0]$ is defined as the length unit (the International Prototype of the Kilogram is another example of such a primary standard).

In this case, metrological traceability requires the empirical accessibility and stability of s_0. In the simplest case in which the objects a and b can be actually compared to s_0 by length, Premise 3 amounts to a guarantee that (a) s_0 was used in both comparisons, and (b) the length $L[s_0]$ did not change while producing the two results in Premises 1 and 2. More generally, the traditional solution has been to identify or construct a sequence of measurement standards, $\langle s_1, s_2, ...\rangle$, such that, for $i \geq 0$, s_{i+1} is accessible by s_i and $L[s_{i+1}]$ is guaranteed to be empirically indistinguishable[7] from $L[s_i]$:

$$L[s_{i+1}] \approx L[s_i]$$

Each of these operations is called the *calibration* of the i+1th standard, and the whole sequence is called a *metrological traceability chain* (JCGM, 2012: 2.42), which disseminates the unit $L[s_0]$ so that the traceability of measurement results obtained by comparing the measurand and the length $L[s_n]$ of a *working standard* s_n is guaranteed. The set of all traceability chains rooted on the same primary standard is called a *metrological system*, and is the core organizational structure aimed at making measurement results of physical quantities intersubjective[8].

3.2. The newer solution in physical measurement

The traditional strategy just described has two obvious drawbacks, both related to the fact that the metrological system is rooted on one object, the primary measurement standard. The first

[6] Individual properties are assumed to be entities that exist, and as such are *identified* and not defined (for the same reason for which we can identify, say, a table, but we cannot define a table, but only what a table is, i.e., the concept 'table'). On the other hand, units are individual properties that are identified for a given purpose: to operate as reference quantities. Hence, what is actually defined in a unit definition is not the individual quantity as such, but the mode of identification of the individual quantity chosen as the unit.

[7] In general, such empirical indistinguishability involves some uncertainty, which is added with each calibration (from s_n to s_{n+1}). Thus, calibration uncertainty usually increases along the traceability chain. As this does not affect the arguments presented in this paper, we do not consider it further here.

[8] The structure of metrological systems built for physical properties measured on algebraically weaker scales is basically the same, and only requires more measurement standards to be disseminated and calibrated.

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

is that ownership of a measurement standard becomes a critical source of control over the whole system. The second is that the stability of a metrological system depends on the stability of the primary standard, a condition that for macroscopic physical objects can be guaranteed only to a given extent (for example, as is now well-known, the aforementioned International Prototype of the Kilogram has changed mass by a small but detectable amount since its initial adoption).

Hence, an alternative strategy has been systematically adopted, which involves identifying a reference property as that realized by a phenomenon (or, equivalently, a reference property of a class of objects), in given conditions, assumed to be stable according to the best available relevant theory. Again in the case of length:

$$\mathrm{m} := L[P_C]$$

where P_C is, for example, a beam of light in vacuum for a given time duration, whose length $L[P_C]$ is defined as the length unit.

While in principle the identifying phenomenon is universally accessible, and therefore the working standards could be calibrated against it, in practice its accessibility may be limited (as is indeed the case for the speed of light in vacuum used to define the metre: the experimental equipment to reproduce the phenomenon would not be affordable for most factories needing to calibrate their length measuring instruments). Hence primary standards and traceability chains still remain useful. In this role, a primary standard s_0 is not the object a quantity of which is by definition the unit, but an object that realizes the independent definition of the unit[9]. Such a realization, performed according to a procedure called a "mise en pratique" (www.bipm.org/en/measurement-units/rev-si/mise-en-pratique.html), aims then at guaranteeing the equivalence of the quantity realized by a primary standard to the unit:

$$L[s_0] \approx L[P_C]$$

Thus, multiple primary standards can be identified or constructed, each of them possibly operating as the root of distinct traceability chains, and a new problem arises of comparing the relevant quantities of these primary standards in order to assess their equivalence and therefore to guarantee that through such traceability chains traceability to the same unit is obtained.

Relating to the units of the International System of Units (SI), the organizational solution that has been found for these comparisons is the CIPM Mutual Recognition Arrangement (MRA, www.bipm.org/en/cipm-mra), a framework that requires the National Metrology Institute (NMI) of each Member State to maintain one or more primary standards for each general property (more specifically, quantity) considered in the framework and to demonstrate the equivalence of such standards through their periodic comparison with the standards of the other participating NMIs. For each general property involved in the framework, the outcome of these so-called "key comparisons" is a set of internationally peer-reviewed, recognized, and publicly available documents: each of them states, in particular, the key comparison reference value with its associated uncertainty, and for each NMI, the value of the property of each standard and its declared uncertainty, together with the deviation from the key comparison reference value and the uncertainty in that deviation, i.e., the "degree of equivalence" of the properties of the standards involved in the key comparison.

[9] The transition from the traditional to the newer solution also implies a change in the structure of unit definition: in the traditional solution the unit is defined as a quantity of the primary standard; in the newer solution the unit is defined, in principle, independently of any standard.

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

3.3. Underlying principles

Despite their differences, the solutions just described share three basic principles:

(i) identifying a reference scale and using it are distinct processes;
(ii) measuring instruments are designed and operated under the assumption that the reference scale (the unit, for quantities) for the property they measure is already and independently defined;
(iii) measuring instruments are designed and operated so as to compare the property to be measured and the reference scale (for quantities, the unit and a given set of its multiples): the property to be measured is a property of the object under measurement; the reference scale is either conveyed by some measurement standards (in direct synchronous methods of measurement) or is stored in the instrument via its calibration (in direct asynchronous methods of measurement) (Maul, Mari, Torres Irribarra, & Wilson, 2018).

4. From physical to human measurement

Given our position that measurement results should be context-independent and identically interpretable by different measurers (i.e., they should be intersubjective), the very idea that some psychosocial properties can be measured is critically conditioned by the same constraint. Consider the example of the chess playing ability (*CPA*) of individuals. An inference like the one presented previously should then apply:

Premise 0: chess playing ability can be measured on an ordinal scale
Premise 1: $CPA[a]$ has been measured to be 7 on the scale s_1
Premise 2: $CPA[b]$ has been measured to be 8 on the scale s_2
Premise 3: $s_1 = s_2$
Conclusion: then $CPA[b] > CPA[a]$

Similarly for the example of the reading comprehension ability (*RCA*) of individuals:

Premise 0: reading comprehension ability can be measured on an interval scale[10]
Premise 1: $RCA[a]$ has been measured to be 1.23 times a given unit u_1 above a zero z_1
Premise 2: $RCA[b]$ has been measured to be 2.34 times a given unit u_2 above a zero z_2
Premise 3: $u_1 = u_2 = u$, and $z_1 = z_2$
Conclusion: $RCA[b] = RCA[a] + (2.34 - 1.23)\ u$

The structure of these arguments is identical to the physical measurement examples presented previously. As before, the intersubjectivity of measurement depends on the traceability of results to the same scale (Premise 3). The challenge, then, is to specify how this condition can be guaranteed.

[10] It is rarely possible to identify a natural zero for putative psychosocial quantities. As a consequence, in addition to identifying a unit, it is also necessary to identify an origin (zero) for the scale (that is, in Stevens' terms, the resulting scale is interval, but not ratio, analogous to Celsius and Fahrenheit scales for temperature). A variety of conventions exist for doing this, such as identifying the origin in terms of the average of a specified population at a specified time, or the average difficulty of a given set of items.

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

4.1. The analogue of the "traditional solution" in ordinal measurement of human properties

Given that in the human sciences the objects that bear properties are often human beings, a solution parallel to what was previously described as the "traditional solution" in physical measurement would require that a person (or set of people, in the case of an ordinal property) be available as a reference, to which other persons could be compared in terms of the relevant property. This is in fact not so different from what could be considered some of the earliest forms of measurement of human properties, as when the skill or ability of individuals is assessed via direct competition with other individuals: in the case of an ordinal property, such as chess playing ability as considered previously, one could imagine, for example, that a given set of persons with progressively higher levels of skill is chosen as the set of the standard objects, and that the skill of an individual is assessed by putting that individual into competition with these progressively-more-skilled "standard" individuals to determine where in the sequence the new individual could be located, a solution structurally similar to the application of Mohs scale for measuring the hardness of a sample mineral[11].

As before, though, such a solution depends on both the accessibility of the individuals chosen as references and on their stability. The direct accessibility of such individuals could only plausibly be guaranteed in a very small community; in principle, traceability chains with sequences of such standard individuals could be set up in a manner parallel to that previously described for physical properties. But even if the first condition were solved by means of such chains, the second condition is potentially even more problematic, for the basic reason that human beings learn, and more generally change over time in a myriad of ways. If the number of categories used were small and the difference between them very significant – for example, a novice, expert, and master chess player – perhaps one could plausibly expect the ordering to remain invariant over a relevant time period, but at scale this condition would also be virtually impossible to guarantee. As such, the need to identify references in terms of classes of objects (specifically, classes of persons) rather than actual objects (persons) is arguably even more pressing in the measurement of psychosocial properties.

4.2. The analogue of the "newer solution" in ordinal measurement of human properties

The "newer solution" in physical measurement, as previously described, involves identifying a reference property as that realized by a phenomenon or a class of objects, in given conditions, assumed to be stable according to the best available relevant theory. For most applications in the human sciences, it is not obvious that there are any known natural phenomena that can serve in this role. However, a solution with a similar structure may still be possible. Returning to the example of chess playing ability discussed in the previous section, we can imagine that the behavior of the aforementioned set of progressively more-skilled actual chess players could be captured, encoded, stored, and made available by means of artificial intelligence tools, as a set of prototypical virtual chess players; the chess playing ability of an individual could then be assessed using the same procedure previously described, by competing with a

[11] We ignore here measurement uncertainty (in particular, instrumental uncertainty), but in practice it would be necessary to find a way to deal with the possibility of influence properties acting on the outcome of specific games of chess.

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

computer program that instantiates these virtual players. In principle, this solves the two major problems of empirical accessibility and stability: first, given that these virtual players are informational entities rather than physical entities, they can be easily disseminated, and second, algorithms are by definition stable[12]. Thus, Premise 3 can be satisfied for anyone using the standard set of virtual chess players in the manner described, and a conclusion like $CPA[b] > CPA[a]$ can indeed be reached from Premises 1 and 2. As can be seen, the solution described in this example is not structurally different from the physical example of hardness described earlier, despite the differences in subject matter.

4.3. The analogue of the "newer solution" in quantitative measurement of human properties

While the solution just described is in principle satisfactory for ordinal cases (without significant instrumental uncertainty), in the context of many areas of the human sciences, such as in psychology and education, it is more common for putative measurement results to be modeled as quantities, and reported on interval scales; also, it is not common for measuring instruments to operate by comparing a person with another person, virtual or otherwise. Commonly, in contexts such as psychology and education, a measurement instrument consists of a set of specific, constrained challenges, prompts, or questions (generically termed "items"), to which individuals are required to react in some fashion. For many kinds of tests it is common for these items to be presented with a fixed number of possible responses (as in the canonical cases of multiple-choice test questions on an academic test, or rating scales on surveys) which have been predetermined to be correct or incorrect (or, more generally, to be indicative of higher or lower knowledge or skill, or more or less severe attitudes, etc.), but more open-ended options are used as well, such as a challenge in which a person must produce a piece of writing or other performance which is then evaluated. The question relevant to the central argument of this paper could then be formulated in terms of seeking a justification for Premise 3 in the arguments previously developed, for example for reading comprehension ability: if the *RCA* of two different individuals in two different places and times have been measured to be 1.23 and 2.34 on scales s_1 and s_2, respectively, is it possible to infer something about the relationship between the *RCA* of these two individuals?

A possible solution to this problem comes in the form of an extension of the logic of identifying virtual reference persons developed in the previous sections. To explore this solution, let us suppose that a theory is developed in which *RCA* is assumed to be a quantity, and that a set of items is developed based on this theory. Suppose further that item-response data from the administration of these items to individuals in the relevant population is dependably found to fit the Rasch model[13], which posits a stochastic relationship between a property (more specifically, quantity) of persons and success or failure on each of the items, depending on their severity (or difficulty). A generic version of the Rasch model can be written as:

[12] It could be argued that there are exceptions to this statement, for example neural networks, which might be considered to be (based on) algorithms that are not necessarily stable, due to changes in parameters over time. But even in such cases, the algorithm can still be in principle completely characterized.

[13] For an introduction to the Rasch model in the perspective of physical measurement and an expanded account of approaches to measurement using the Rasch model and how they relate to physical measurement, see, e.g., Mari & Wilson (2014). For more general perspectives on the role of the units in Rasch models, see, e.g., Humphry (2011) and Humphry & Andrich (2008).

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

$$\text{log-odds}(X_{n,i} = 1) = \theta_n - \delta_i$$

where θ_n is the considered property of the *n*th person, δ_i is the severity of the *i*th item (which can be defined as the value of θ above which an individual is more likely than not to succeed on the item), and $X_{n,i} = 1$ (= 0) is the event that the *n*th person succeeds (fails) in responding to the *i*th item[14]. In the specific case of *RCA*, the equation could then be rewritten as:

$$\text{log-odds}(X_{a,b} = 1) = RCA[a] - RCA[b]$$

where $RCA[a]$ is the reading comprehension ability of person *a*, $RCA[b]$ is the reading comprehension ability required for the successful completion of item *b*, and their difference establishes the log of the odds of success (the equation here is presented without subscripts, but in practice each person would come into contact with multiple items, and vice versa).

The analogy between the cases of chess playing ability described in the previous section and of reading comprehension ability is manifest: in the case of *CPA*, the comparison is between the object under measurement (typically a human chess player) and a virtual chess player; in the case of *RCA*, the comparison is between the object under measurement (typically a human respondent) and a test item, which is interpreted as a tool that instantiates a given level of *RCA*. In other words, while a virtual chess player is an *active* implementation of a standard of *CPA*, a test item is a *passive* implementation of a standard of *RCA*.

To define an interval scale, any individual *RCA* could be defined as the zero, and a unit can then be defined as the distance between any two arbitrarily-chosen *RCA*s. This opens up a range of possibilities for the definition of the zero and the unit, which could be tailored to the purpose and context of the test. For an educational test, for example, the zero might be defined in terms of the average *RCA* for students at a given grade level, and the unit could be defined as the distance between *RCA*s that correspond to the ability to meet specified performance standards with specified probabilities, or that are reflective of typical students at particular grade levels in a specified population (cf. Briggs, 2018). Given the symmetry of the model, the zero could also be defined as the average difficulty of a given set of items deemed appropriate for a given grade level, and the unit could be defined in terms of the difference between the *RCA* required by two items representative of what a student should be able to successfully complete at different points in their education.

If a unit and zero were defined in such a manner for *RCA*, they would together define an interval scale, which we could call s. A measurement result for person *a* could be reported as:

$$RCA[a] = 1.23\ °\text{s}$$

where °s is read as "degrees in scale s", and thus references the scale, including both the zero and the unit, analogously to how °C references to the Celsius scale, which requires the definition of both a zero and a unit of temperature. Premise 3 could then be satisfied for comparisons of persons whose *RCA* had been measured by the particular instrument just described. Thus, the solution described so far has a clear and important limitation, which is that the scale is defined with reference to a specific set of items. This would be analogous to a situation in which the

[14] Extensions of the model are also available for more complex cases, e.g., when item responses have scoring options beyond simply being correct or incorrect (see, e.g., van der Linden, 2018).

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

reference temperature were identified as properties of a given class of thermometers, rather than of objects or phenomena having a given, supposedly constant, temperature. Such a strategy would then inherit many of the established limitations of operationalism in general (see, e.g., Chang, 2009; Maul, Torres Irribarra, & Wilson, 2016), and in particular would have the consequence that the scale would not be of *RCA per se*, but of *RCA* as measured by a specified instrument.

If a sufficiently large set of items were developed that all measured the same property and were all dependably found to fit the Rasch model, as described previously, the *RCA* of any new individual could be measured via administration of any sub-sample of items from this collection of items, usually called an "item bank". In this way, independent instruments (that is, tests with no overlapping questions) could be created, and results from these instruments would be traceable to a common set of reference properties. Premise 3 could then be satisfied for comparisons of persons whose *RCA* had been measured using any subset of the items in the item bank: the single-instrument dependence of the first solution has thus been overcome.

If item-response data from other instruments designed to measure *RCA* could be shown to jointly fit the Rasch model along with the previously described item bank, this solution could be extended further, and Premise 3 could be satisfied for comparisons of persons whose *RCA* had been measured by any of these instruments.

This can be presented in terms of a three-step progression, from a reference scale operationally defined in terms of a single instrument to one whose definition is in principle independent of any given instrument:

1) a reference scale is identified in terms of a given instrument;
2) a reference scale is identified in terms of a given item bank;
3) a reference scale is identified in terms of multiple instruments,

where the third step has the same structural role that the CIPM MRA, as described in section 3.2, has for the units of the SI.

5. Conclusions

Making measurement results traceable to the same reference scale allows for intersubjective understanding of their meaning, a key feature of measurement of both physical and psychosocial properties. On the other hand, the importance of this challenge does not seem to be widely recognized as such in the literature on measurement in the human sciences. Instead, in many areas of research it is common for instruments to be developed locally and ad hoc, even when other instruments have been developed to measure the same property elsewhere; thus, at best, a reference scale might be established for that instrument (step 1, above: the idiosyncratic situation called "pre-measurement" by Frigerio, Giordani, & Mari, 2010). The development of item banks are common in some settings such as large-scale educational testing, in which there is an incentive to develop multiple tests with non-overlapping questions that can be compared to one another; in such situations, at best, a reference scale might be established for that item bank (step 2, above). However, in both cases the argument developed previously for reading comprehension ability (and analogously for other properties) would have to be modified, particularly with regards to Premises 1 and 2:

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

Premise 0: reading comprehension ability can be measured on an interval scale
Premise 1: $RCA[a]$ has been measured to be 1.23 using instrument or item bank b_1 with scale s_1
Premise 2: $RCA[a]$ has been measured to be 2.34 using instrument or item bank b_2 with scale s_2
Premise 3: $s_1 = s_2 = s$
Conclusion: $RCA[b] = RCA[a] + (2.34 - 1.23)$ on scale s

This formulation highlights the significant fact that, in the human sciences, reference scales (including units) often depend on instruments, whereas this is not the case in the physical sciences (see also Fisher, 2007). Further, for the most part, in the human sciences one does not find serious and sustained efforts to move towards instrument-independence (step 3, above). However, as we hope has been shown, this is not due to any structural difference between the two contexts. Indeed, traceability to a reference scale can, at least in principle, be structurally guaranteed for both physical and non-physical properties.

One reason for the apparent resistance in the human sciences to endeavoring to make reference standards instrument-independent may simply be the difficulty in meeting the necessary conditions (see also Finkelstein, 2003, 2009). Although the conditions for traceability for interval measurements have been described here in statistical terms (in particular, in terms of fit to the Rasch model), in practice being able to satisfy such statistical criteria generally requires a well-developed theory of the property and its modes of interaction with other phenomena – which, of course, are basic requirements for *objectivity*, the other major dimension of the quality of measurement results (see, e.g., Mari, 2003). This is no easy task, especially for properties as complex and dynamic as those of interest to human scientists: in particular, these properties are often defined in such a way as to be indexed to particular socio-historical conditions, meaning that as society changes, the very definition of a property may also change. For example, the definition of what reading comprehension ability is has itself changed significantly over the past few decades (e.g., Pearson & Hamm, 2005), as the kinds of texts and materials with which we engage have evolved. Stability of measurement standards, as discussed in this paper, requires a more fundamental *definitional* stability, showing again that measurement quality ultimately requires both intersubjectivity and objectivity.

The structure of the inference we have discussed here – from Premises 0-3 to the Conclusion – throws some light on the very concept of intersubjectivity of measurement. Premises 1 and 2 state that the two properties under consideration have been *measured*, rather than generically evaluated. Were Premise 3 false, the conclusion would then be that such measurements have poor intersubjectivity, not that they are not actually measurements. That is, Premise 3 is required for inferring the Conclusion but is not necessary for measurement. This presentation is consistent with the historical situation of different communities measuring (for example) lengths on different scales: even if the relation between such scales were not known, the claim that in each of the communities lengths were measured would be maintained. On the other hand, Premise 3 admits a continuum of conditions, from complete non-intersubjectivity to complete intersubjectivity, of which the three-step progression introduced above identifies some exemplary cases, from the private scale of an uncalibrated measuring instrument to a system of scales that are universally accepted and provide in principle global metrological traceability, as it is the case of the International System of Units.

Despite the special challenges faced by human scientists and the obvious differences between physical and psychosocial properties, there are structural means to guarantee objectivity

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

and intersubjectivity in both settings. By embedding these requirements in a single and consistent framework, a unified conception of measurement across the sciences can be developed.

References

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

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

Briggs, D. (2018). Tolerating approximate answers about student learning. Paper presented at Oxford University Center for Educational Assessment (OUCEA) Annual Lecture 2018.

Chang, Hasok, "Operationalism", The Stanford Encyclopedia of Philosophy (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/operationalism/>.

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

Finkelstein, L. (2009). Widely-defined measurement: An analysis of challenges. *Measurement, 42*, 1270-1277.

Fisher, W. M. (2007). Living capital metrics. *Rasch Measurement Transactions, 21*(1), 1092-1093.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*, 123–149.

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

Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement*, *9*(3), 249-264.

preprint of a paper under publication in:
Measurement, doi:10.1016/j.measurement.2018.08.068

JCGM (2012). *International Vocabulary of Metrology (VIM) – Basic and General Concepts and Associated Terms (2008 edition with minor corrections)*, Joint Committee for Guides in Metrology, http://www.bipm.org/en/publications/guides/vim.html.

Krantz, D. H., Suppes, P., & Luce, R. D. (1971, 1989, 1990). *Foundations of measurement, volumes I, II, and III*. New York: Academic Press.

Mari, L. (2003). Epistemology of measurement. *Measurement, 34*, 17-30.

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

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

Maul, A., Mari, L., Torres Irribarra, D., & Wilson, M. (2018). The quality of measurement results from a structural perspective. *Measurement, 116*, 611-620.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. Paris (Ed.), *Children's reading comprehension and assessment*, ch. 2.

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

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

Van der Linden, W. J. (2018). *Handbook of item response theory, three volume set*. Chapman and Hall/CRC.