A comparison of measurement concepts across physical science and social science domains: instrument design, calibration, and measurement

M Wilson1#, L Mari2*, A Maur3, D Torres Irrbarra1
1 Graduate School of Education, University of California, Berkeley, CA, USA
2 School of Engineering, Università Cattaneo – LIUC, Castellanza (VA), Italy
3 School of Education, University of Colorado, Boulder, CO, USA

E-mail: MarkW@berkeley.edu

Abstract. Is there a framework common to measurement in both physical and social sciences? The answer to this question would determine to an important extent the possibility of building a shared measurement-related body of knowledge across these traditionally separate domains. In this paper, we outline a framework of the processes involved in the construction and use of measures that includes instrument design, instrument calibration, and ultimately measurement using the instrument. A comparison of these steps across the two domains reveals both (a) formal parallelism, and (b) important differences in the way calibration is intended and implemented. We examine the similarities and differences to determine whether this is a case of irreducible difference, or whether the similarities are such that measurement in the two domains can be viewed within a single conceptualization.

1. Introduction
Measurement results are expected to convey information on the measurand in terms of the adopted scale (e.g., the numerical quantity value reported in the measurement result, the ratio measurand/unit), and to be independent of the adopted measuring instrument and of quantities other than the measurand.

The exclusive dependence of measurement results on the measurand would guarantee [a] the objectivity of measurement information, i.e., its object-relatedness, and the reference to a universally accessible scale would guarantee [b] the intersubjectivity of measurement information, i.e., its subject-independence, where object-relatedness and subject-independence can be considered characterizing features of measurement itself [1], [2]. How can they be obtained?

The answer to this question has nothing to do with the (possibly quantitative) form of the measurement results [3], but is grounded instead on the experimental structure of measurement systems: it is the suitable (i) design, (ii) calibration, and (iii) operation of such systems that gives measurement results the features of object-relatedness and subject-independence to an acceptable degree.

(i) Design. In the design stage the two basic requirements are:

# 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.
[a] measuring instruments are designed so to have a stable behavior, sensitive to the measurand and insensitive to external influences (‘influence quantities’), thus aiming at obtaining object-related information;  
[b] measurement standards are designed to allow the realization of scales and units that are stable and socially widespread, thus aiming at obtaining subject-independent information.  
(ii) Calibration. The outcome of the design stage is generally a class of measuring instruments or measurement standards, which must be then realized, i.e., instanced in specific devices that can be operated here-and-now. This requires establishing the values in the here-and-now conditions of the parameters left unspecified in the design, an outcome obtained by the calibration stage. Calibration is a process aimed at: 
[a] removing instrument influence from the measurement results, and thus obtaining object-related information;  
[b] making measurement results traceable to a given scale/unit, and thus obtaining subject-independent information.  
(iii) Measuring. Finally, a calibrated measuring system can be operated in the measurement stage, where the operation can proceed: 
[a] without interference from outside influences, and  
[b] with consistency across subjects.  
This framework justifies the structure of the meta-process: design first, then calibrate, and finally measure. The question we are addressing in this paper is the following: Is this structure common to measurement in both physical and social sciences? A positive answer to this question would seem to be a precondition for building a shared measurement-related body of knowledge across these different domains.  
Some historical background in the social sciences on this can be gathered by considering the views of O.D. Duncan: “Let us note the clear separation between the two tasks: first, to scale the statements, or to infer their location on a linear attitude continuum defined in terms of the polar contrast of “strongly affirmative” with “strongly negative”; and, second, to place respondents on that same attitude continuum insofar as their locations on it can be inferred from the selection of items they choose to endorse. Not only are the two problems logically distinct and operationally separated, but also the former – the scaling of statements – precedes and is presupposed by the latter.” [4: p.174].  
Despite the differences in language and content (the description above is about Thurstone’s scaling, a basic measurement method in the social sciences [5]), the reference is manifestly to a situation that is usual to physical measurement too: first, the measuring instrument is designed and calibrated; second, it is used in measurement, where calibration and measurement are indeed, in Duncan’s words, “logically distinct and operationally separated” processes and “the former precedes and is presupposed by the latter”. On the other hand, Duncan continued: “By contrast, psychometric or “test theory” methods of attitude measurement either attempt to accomplish the two tasks simultaneously and jointly or even to bypass the scaling of items altogether.” [4: p.174].  
Here the pivotal role of calibration is under discussion: is it really possible to measure and to calibrate at the same time, or to measure by means of an uncalibrated instrument? Duncan has demonstrated that at least some social science measurement approaches do not seem to conform to the 3-stage system described above. Now, our task is to investigate whether this is a case of irreducible difference (i.e., the same term, “measurement”, is used with two irreconcilably different meanings), or whether the differences may instead be merely due to differences in commonly-used terminology and conventional practices between the two domains.  
2. A black box approach to measuring instrument design, calibration, and measurement  
With the aim of introducing what may constitute the simplest sharable basis for a measurement-related conceptual framework, we propose a purely structural view of the meta-process under consideration as follows (an even simpler description would be about measurement by direct comparison, as performed by a ruler or an equal arm balance, but this case is much less interesting for our purposes). The reader
can interpret what follows as the design, calibration, and operation process of, say, a simple thermometer.

1. One observes that a property \( Y \) of an object \( b \) changes (say, the length \( Y(b) \) of \( b \)). From further observations one hypothesizes that the changes of \( Y(b) \) are systematically and regularly caused by the changes of a property \( X \) of another object \( a \) that somehow interacts with \( b \) (say, the temperature \( X(a) \) of \( a \), being it possibly the environment surrounding \( b \)). "Systematically and regularly" only means here that, ceteris paribus, if in different times, \( t_1 \) and \( t_2 \), one observes \( Y(b(t_1)) \) is [not equal to] \( Y(b(t_2)) \) (the length of \( b \) is [not] the same at the times \( t_1 \) and \( t_2 \), then she hypothesizes that also \( X(a(t_1)) \) is [not] equal to \( X(a(t_2)) \) (the temperature of \( a \) is [not] the same at the times \( t_1 \) and \( t_2 \)). In this situation one can write \( F(X(a)) \) to denote the property \( Y(b) \) caused by \( X(a) \), where then \( F \) is a causation mapping that describes the transduction \( X \rightarrow Y \) performed by \( b \). One also observes that the same behavior of \( b \) is obtained by its interaction with a whole set of objects \( \{a\} \), i.e., all \( a \) have the property \( X \) and all \( X(a) \) systematically and regularly produce corresponding changes of \( Y(b) \).

2. A socially widespread enough and stable enough object \( a^* \) in \( \{a\} \) is available, so that \( F(X(a^*)) \) can be easily obtained and generally \( X(a^*(t_1)) = X(a^*(t_2)) \) (how can the relationship 'having the same temperature' be determined without recurring to \( b \)? let us omit the discussion of this subject here).

3. Given the object \( a^* \) the relation 'having the same \( X \) as \( a^* \)' (having the same temperature as \( a^* \)) can be experimentally determined by means of the object \( b \), according to the procedure:

   - \( P \) for each candidate object \( a \), in \( \{a\} \), \( a \) has the same \( X \) as \( a^* \) if \( F(X(a)) = F(X(a^*)) \) (note that this procedure does not require the comparison, nor the comparability, of \( X(a) \) and \( X(a^*) \)).

   The information that \( a \) has the same \( X \) as \( a^* \) can be written \( X(a) = X\text{-of-}a^* \) for short.

   Hence:
   - \( a \) is the object under measurement;
   - \( a^* \) is an object used as a measurement standard;
   - \( b \) is used as a measuring instrument;
   - \( X \) is the property subject to measurement;
   - \( Y \) is the instrument indication;
   - \( X\text{-of-}a^* \) is the value realized by the measurement standard,

   and:
   - Stage 1 corresponds to the design of the measuring instrument \( b \);
   - Stage 2 corresponds to the selection of a measurement standard and the (first) calibration of \( b \);
   - Stage 3 corresponds to a measurement performed by means of \( b \) according to the procedure \( P \).

   Note also that:
   - only the causal relationship between \( X \) and \( Y \), as realized by \( b \), is assumed: no known mathematical functions or models are required;
   - despite this simplicity, model-ladenness is unavoidable, at least in the assumptions of ceteris paribus comparison (i.e., all influence properties did not change), and stability of the standard.

   Unfortunately, something else must be added to this description.

4. The transduction behavior \( X \rightarrow Y \) of \( b \) could be recognized to be not perfectly stable, i.e., \( F(X(a^*(t_1))) = F(X(a^*(t_2))) \) even if \( X(a^*(t_1)) = X(a^*(t_2)) \) (the caused lengths are different even if the causing temperature remains the same). This shows that in general the mapping \( F \) could be (for example) time dependent in its turn (where the causes of this time dependence might be in the failure of the ceteris paribus hypothesis, but again we will not develop this subject here). Hence \( F(t, x) \) must be established so that from \( Y(b(t)) = F(t, X(a(t))) \) (the length of \( b \) obtained by its interaction with the temperature of \( a \) at the time \( t \)) one can obtain whether \( X(a(t)) = X\text{-of-}a^* \) or \( X(a(t)) \neq X\text{-of-}a^* \). The idea is, trivially, of identifying \( F(t, X(a^*(t'))) \) at a time \( t' \) "close enough" to the measurement time \( t \), so that one can suppose that the following time-dependent version of the procedure \( P \):

   \[ P' \] for each candidate object \( a \), in \( \{a\} \), \( a \) has the same \( X \) as \( a^* \) if \( F(t, X(a(t))) = F(t', X(a^*(t'))) \)

   can be applied. This stage corresponds then to a (re) calibration of \( b \).
This shows the most fundamental reason why instrument calibration and measurement are (complementary but) distinct processes: the former operates on an object – a measurement standard – whose quantity under consideration is assumed to be known, whereas the latter operates on an object – the object under measurement – whose quantity under consideration is assumed to be unknown and in fact looked for by means of measurement itself.

3. Social science measurements: the case of norm-referenced measurement

Now, the question can be asked: does the black box approach described in the previous Section apply to (at least some) social science measurements? As the quotation from Duncan above shows, it is not straightforwardly clear that the response is always “yes”. In fact, it must be said that social science measurement does not usually allow for the concept of a measurement standard (i.e., the object \(a^*\)) in the straightforward way that it is used above. There are, for example, no actual persons who are generally agreed to be the measurement standards of 100 IQ (or 115, 130, etc.), and this is also true in other social science applications such as the measurement of attitudes, and academic knowledge and skills. Instead, the most common form of standardization is based on a so-called “norm-referenced” perspective where, say, the mean response in a population is labeled with a certain numeral (e.g., 100 in the case of IQ tests), and other points are given labels depending on their location in the frequency distribution of the responses (such as 115 for 1.0 standard deviations above the mean, 101 for 1/15th of a standard deviation above the mean, etc.). Clearly, this technique depends on not only the instrument as it is usually conceptualized (usually as a set of items to which the persons respond in some way, which are then categorized according to a particular theory), but also on the sample of persons that were used to calculate the statistics on which the labels were based (i.e., the mean and standard deviation of the population in the IQ example). Thus, if one extends the concept of an instrument to include the particular sample to which it was administered (or perhaps more broadly, to a definition of a “sample representative of a population”), then the status of this as a measurement standard can be (at least abstractly) claimed. In fact, there are historical cases where this dependence on a specific sample has been maintained over extended periods of time (one example is the commonly used scale of scores for the SAT, which were originally established on a particular sample of college entrance candidates in the 1940s, and then used as a consistent framework through the 1990’s, although the link has been subsequently abandoned). The manner in which this historical link is maintained (or is not maintained) can have a range of interesting consequences. For example, Flynn’s [6] showed that, according to the results of typical IQ tests, the IQ of the populations of many countries has risen considerably throughout the 20th century, yet this phenomenon was not generally apprehended by IQ researchers during the 20th century precisely because they continually updated their IQ scales to match the current population, and the norm-referenced standardization described above concealed the change.

The details of how the norm-referenced approach can be conceptualized as a calibration are given in [6]. The essential insight is to see that the specific set of response vectors that correspond to an IQ score of, say, 100, are indeed the measurement standards here. The correspondence of this example to the black box approach given above is given by the following:

- \(a\) is the object under measurement (a person);
- \(X(a)\) is the intelligence of \(a\);
- \(b\) is the IQ test (including both content and technical specifications);
- \(Y(b)\) is the vector of responses to \(b\);
- \(y\) is the indication value, the so-called “score” of \(Y(b)\), and for each measurement standard:
- \(a^*\) is the set of persons whose response vectors have the same score as that which has been assigned for that measurement standard (e.g., the set of persons who have a score that is sufficiently close to the mean).

Now to make this a bit more concrete, take a (non-typical, but small enough to be clear) case where there are 10 dichotomous items in an IQ test, and a score of, say, 3 is found to be the score closest to
the mean, and hence given a numeral of 100. In this case, there are $2^{10} = 1024$ distinct response vectors, of which 252 correspond to the measurement standard (the mean), labeled as 100.

We can then see how the three steps of the instrument design, calibration, and operation are carried out in this situation. First, the instrument development would be constituted by the selection of a set of items whose responses were believed to be related (positively) to a person’s intelligence. Second, the calibration would consist of gathering the item response vectors of a sample of persons from a suitable population, and calculating the required statistics for the calibration (such as the item difficulties, the mean person score, etc.), and the assignment of the response vectors to the labeled measurement standards (such as the mean, with a score of 100). Third, the operation of the instrument occurs when a person now takes the IQ test, and thus generates a response vector, the indication value of which is compared to the various measurement standards (e.g., if her response vector has a score greater than the mean, she will have a label greater than 100, etc.). Thus, in a formal sense, this procedure of norm-referenced calibration does indeed have the same structure as the black box approach presented above.

However, what one observes from this situation is that these measurement standards are not very satisfactory: as noted above regarding historical examples, they are sensitive to the particular sample that was used to calculate the mean and other statistics used, which prima facie seems to suggest that the results obtained will not be independent of the particular instrument used, and will thus not be objective. In addition, a second disadvantage is that the connection from the basic and interpretable parts of the instrument (i.e., the items and the coded responses) to the standard (i.e., the specific set of response vector that corresponds to a particular norm-referenced measurement standard) is very complex, and hence difficult to use for any interpretational purpose. To see this, consider the example mentioned where an IQ score of 100 corresponds to 252 distinct response vectors! It is not at all clear how one would make an interpretation of such a score based on the individual 10 items and their respective scoring codes, which were the initial constructive pieces of the test. Of course, with more items (as there typically are), the number of distinct response vectors becomes much higher, and even more so when polytomous scores allowed on the individual items. Thus, these measurement standards are not of much practicable use, and hence, it is our interpretation that it is for this reason that their conceptualization as “measurement standards” has been ignored by the very psychologists and others who have implicitly used those standards so frequently.

With this in mind, one can interpret the comment of Duncan that “psychometric or ‘test theory’ methods of attitude measurement either attempt to accomplish the two tasks simultaneously and jointly or even to bypass the scaling of items altogether” as being both incorrect and insightful at the same time. He has not recognized the distinct parts of calibration and measurement in the standard norm-referenced approach, and has not acknowledged the role of the items in the creation of the measurement standards. But, he has also correctly, and insightfully, observed that the calibration itself is not very satisfactory, as it renders the results dependent on the particular sample of persons that is used, and that the (crucial) role of the items is rendered all but invisible in the process.

The question that remains is whether this unfortunate situation can be remedied. That is, can one replace the problematical technique of norm-referenced calibration with a better one--one that reduces (or, ultimately, eliminates) the dependence of the measurements on the sample of persons, and also makes the role of the items more transparent in the calibration? In pursuit of this question, Wilson [7] has described several of the major approaches to measurement that have been developed in the last century, including classical test theory (equivalent to the norm-referenced approach described above), Guttman’s scalogram technique [8], Rasch scaling [9], [10], and Construct Modeling [11], and shown how these last three can each be seen as making improvements on the preceding one. We will not repeat that account here, but it will suffice to say that the improvements have not led to the elimination of all problems, but that they are important for the future of measurement in the social sciences, and also form a basis for establishing the structural equivalence of measurement in the physical and social sciences.
4. Conclusion
In this paper we have sketched a comparison between formalized versions of measurement in physical and social science measurement: in sum, they can be seen to have important similarities. We have illustrated this for one of the major approaches to social science measurement, (norm-referenced measurement), noting some points where the comparison seems somewhat tenuous, or, at least, not straightforward. In next steps, we will add a comparison to a second approach to social science measurement (construct modeling), which we see as being closer in its logic to physical science measurement, though differences still remain.

5. References
[1] Mari L 2000 Beyond the representational viewpoint: a new formalization of measurement Measurement 27 71-84
[2] Mari L, Carbone P and Petri D 2012 Measurement fundamentals: a pragmatic view IEEE Trans. Instr. Meas. 61(8) 2107-2115
[3] Mari L, Maul A, Torres Irribarra D and Wilson M 2013 Quantification is neither necessary nor sufficient for measurement IOP J. Phys.: Conf. Ser. 459 012007
[4] Duncan OD 1984 Notes on social measurement. Historical and critical (New York: Russell Sage Foundation)
[5] Thurstone LL 1927 The method of paired comparisons for social values Journal of Abnormal and Social Psychology 21 384-400
[6] Flynn JR 1987 Massive IQ gains in 14 nations: What IQ tests really measure Psychological Bulletin 101 171-191
[7] Wilson M 2013 Using the concept of a measurement system to characterize measurement models used in psychometrics Measurement 46(9) 3766-3774
[8] Guttman LA 1950 The basis for scalogram analysis Studies in Social Psychology in World War Two ed SA Stouffer, LA Guttman, FA Suchman, PF Lazarsfeld, SA Star and JA Clausen (Princeton: Princeton University Press) vol 4 Measurement and prediction pp 60-90
[9] Rasch G 1960 / 1980 Probabilistic models for some intelligence and attainment tests (Chicago: University of Chicago Press)
[10] Wright B and Stone M 1979 Best Test Design (Chicago: MESA Press)
[11] Wilson M 2005 Constructing measures: An item response modeling approach (Mahwah, NJ: Erlbaum)