Preliminary analysis of validation of measurement in soft systems

L Mari, E Ugazio
Università Cattaneo – LIUC, C.so Matteotti, 22 – 21053 Castellanza (Va), Italy
E-mail: lmari@liuc.it

Abstract. Soft metrology, the science of measurement applied to non-physical properties, is becoming increasingly important in many application fields, but the validation of measurement for soft properties is a topic still worth of exploration. This paper aims at introducing some aspects of a general framework to interpret such topic and at showing how measurement-related models can be exploited as validation tools.

1. Introduction

Measurement has become a paradigm of information acquisition process building on the lessons learned on physical quantities, for which the experimental interaction with the system under measurement is performed by means of physical devices transducing quantity-related signals to indications according to a process whose objectivity and inter-subjectivity can be, at least in principle, assessed [1]. As for non-physical properties ("soft" properties for short henceforth, including in particular properties of social systems such as organizational performance and perceived quality of service – in this paper we will not emphasize the distinction between ‘property’ and ‘quantity’, and therefore we will always refer to the former for the sake of generality) this strategy does not apply, their measurability is a critical problem [2, 3] and its solution has been looked for according to different standpoints, even if not simply a priori refused [4]. In particular, the so-called representational theories of measurement were developed precisely to yield a formal criterion of measurability on the basis of purely structural conditions – the known requirement of (homo)morphism from an empirical relational system to a symbolic / numerical one – thus allowing their application independently of the availability of physical transducers and therefore also for soft properties [5].

Some recent revised versions of the representational standpoint, keeping into account aspects of the measurement process such as the calibration of the measuring system and the traceability chain to standards, interpret measurability in the same direction, and basically aim at detailing the conditions for objectivity and inter-subjectivity of measurement in purely functional way, so to give guidelines for measurability that can be applied also when the measurand is a soft property [6]. On the other hand, the most critical problem of soft properties as measurands is sometimes not measurability. Indeed, the definition of a measuring system and a measurement procedure when the target is to acquire information on a property such as, e.g. customer satisfaction could be more or less straightforward.
The critical problem here is instead of obtaining an acceptable guarantee that the measurement data can be interpreted as actually referred to the intended property. A typical example is about Intelligence Quotient: the known troubles about IQ tests are not related to their objectivity or inter-subjectivity, but on the claim that those tests actually measure intelligence instead of something else. What is relevant here is therefore the validity of measurement and its results, a problem that arises as far as a non strictly operational standpoint is assumed: were a property by definition what a measuring system measures and nothing else, each property would be, in a sense, self-validated. On the other hand, for measurement results to be useful as inputs to a decision making process measurands have to be characterized in the semantics of that process, not the data acquisition one. This generally implies that they should be interpreted in the broader context of a predefined “underlying model” of the property under consideration [7]. Hence, the validation problem arises.

2. Analysis of validation

We claim that one of the most important, although still largely implicit, contributions of the International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM3) [8] is the acknowledgment that while measurement is an experimental process some conceptual modeling activity is required for its results to be interpreted and properly used. For example, def 2.1 note 3 states that “measurement presupposes a description of the quantity [to which the obtained quantity values “can reasonably be attributed”] commensurate with the intended use of a measurement result”. The role of “intentions” (i.e., the reasons of “the intended use”) is also emphasized in the very definition of ‘measurand’, “quantity intended to be measured”, particularly if compared with the corresponding definition of the VIM2, “quantity subject to measurement” (see def. 2.3 note 2), of operational flavor.

This holds for any measurand and any measurement, but it is specifically critical in the case of soft properties, for which physical transducers as data acquisition devices cannot be exploited and therefore the meaningfulness itself of the obtained results depends on interpretive models, that become pivotal for the validation of measurement and its results. Because its complexity, the issue can be dealt with according to a divide et impera (“divide and conquer”) strategy, in reference to the structure of the measurement process, interpreted according to a black box (meta-)model and therefore in terms of inputs – primarily the “quantity being measured” but also, e.g., the influence quantities – that enter a transformation process – the measurement itself – that produces an output – the measurement result (see also [9]).

- The inputs are described by means of a measurand model, whose hypothesized quality is expressed in terms of a definitional uncertainty (see def. 2.27; this is a proper rename of the concept previously dubbed “intrinsic uncertainty” in [10], as there is nothing “intrinsic” in it). Such uncertainty is related to the “amount of detail in the definition of the measurand” (as in def. 2.27), and therefore is derived, in particular, from the identification of the measurand itself in terms of its relations with the influence quantities. In the case of soft properties it is precisely the measurand model that should operate as a bridge between the operational interpretation (that simply assumes the measurand to be what is measured by the given measuring system), that is usually too specific for the decision making process, and the underlying model of the property, that is instead too generic.

- The process is designed and interpreted by means of a measuring system model, whose hypothesized quality is expressed in terms of an instrumental measurement uncertainty (see def. 4.24). A critical task of this model is to describe the interaction of the measuring system with the system under measurement, and therefore to keep into account the possible difference between the ‘quantity intended to be measured’ and the ‘quantity (actually) subject to measurement’. As properly noted by the VIM3, “instrumental measurement uncertainty is obtained through calibration” of the measuring system (see def. 4.24 note 1): this highlights the importance, that is often neglected in measurement of soft systems, of calibrating devices such as, typically, questionnaires.
- Finally, measurement results have to be in their turn expressed according to a model aimed at justifying which (definitional and instrumental measurement) uncertainty components are considered and combined in the measurement uncertainty, and which kind of mathematical expression (e.g., pdf, interval, measured quantity value + combined standard measurement uncertainty) should be chosen as better fitting the needs of the decision making process.

It should be noted that, as defined in the VIM3, the concept of measurement model (see def. 2.48) covers only a part of these models.

3. A case study: questionnaires as measuring instruments

Questionnaires are common devices to acquire information about soft properties. Our aim is to highlight here that a metrological analysis is able to identify different usages of questionnaires, each of them presenting its own issues with respect to validation of measurement (for this analysis, see also [11]). The two basic cases we will consider are related to the evaluation of:

- Case 1: the suitability of a candidate for admission to university;
- Case 2: the quality of a university course, as perceived by one or more students (the so-called “external quality”).

In both Case 1 and Case 2, the information on the measurand (suitability for admission, perceived quality of the course) is acquired by means of a questionnaire, hence interpreted as a measuring instrument. In the simplest version of both cases, a questionnaire is composed of \( n \) closed questions (in Case 1 it would be called a multiple-choice test), and the obtained indication is the set of given answers. The measurement procedure specifies how the questionnaire must be submitted (e.g., with all questions still unchecked, and with instructions concerning the duration of the filling operation and the sources of information that can be possibly used). The validation of measurement is critically related here to the models assumed for both the measurand and the measuring system.

In Case 1 the measuring system coincides with the questionnaire and the system under measurement with the candidate who fills it, while the measurand can be evaluated by simply counting the right answers and then comparing the result with a given threshold, under the supposition that the soft property ‘suitability for admission to university’ is properly sampled by the questions in the multiple-choice test and that such questions weigh equally in this sampling.

| Case 1 |
|---------|
| system under measurement | input quantity | measuring system | processing / calibration | measurand |
| candidate | number of right answers | questionnaire | counting right answers, comparing the result with a threshold | suitability for admission to university |

On the other hand, for many reasons different sets of questions could be used. This shows that a strictly operational approach fails here: ‘suitability for admission to university’ cannot be synonymous of ‘number of right answers’, and it has indeed a more comprehensive meaning, that is implicitly specified in its underlying model. Hence, the problem arises of how the assumption of “proper sampling” for such sets can be validated, i.e., of how well any specific set is able to convey information on the measurand as understood according to its underlying model. This also shows that in this case the problem of validation has to be dealt with by means of calibration: under the hypothesis that a threshold is specified as a minimum number of right answers (instead of, e.g., a given percentile of right answers over the whole population of candidates), is precisely such threshold that sets the “local meaning” of the measurand, and therefore it is responsible of the measurement validity.
Case 2 cannot rely on the simplifying hypothesis that the questions have a right answer, thus preventing the counting of answers as a meaningful operation and making complex, because multidimensional, the set of possible indications. Hence, the processing / calibration stage has the demanding task of mapping such multidimensional space to quality degrees / values. Measurement validation is then critically based on this mapping, that can be even thought of as a partial definition of the very concept of perceived quality for the course under consideration. Furthermore, at least two sub-cases can be envisioned here: both assume that one or more students $s_i$ are requested to fill a questionnaire $q$ about their perception of the quality of a course $c$, and from the answers a quality value $v$ has to be inferred. The two sub-cases are characterized by the role attributed to the students $s_i$, as part of either the system under measurement or the measuring system.

Case 2A hypothesizes that the measuring system is specifically $q$, so that each $s_i$ is considered to be part of the system under measurement together with $c$. This implies that the interaction of $q$ with $c \oplus s_i$, thought of as a direct measurement, can only produce a value $v_i$ to be interpreted as the perceived quality “of $c$ according to $s_i$”. The observation that $v_i \neq v_j$ for $s_i \neq s_j$ simply acknowledges that different measured values can be obtained for different systems under measurement.

Given its nature of information entity, the questionnaire (i.e., the measuring system) is perfectly stable, and then must be calibrated only once: recalibration is indeed required only if the mapping from sets of answers to quality values has to be changed, an operation that actually corresponds to a (partial) redefinition of the measurand.

Finally, the customary option of replicating the application of this procedure to the same course $c$ but with multiple students, $s_1,...,s_n$, with the aim of synthesizing the values $v_1,...,v_n$ and thus producing an information specifically on $c$, corresponds to the different measurement of a different measurand, under the hypothesis that the effects of the presence of $s_i$ in the $n$ systems $c \oplus s_i$ can be progressively reduced if the measurement results obtained from them are appropriately synthesized and $n$ is large enough.

| Case 2A | system under measurement | input quantity | measuring system | processing / calibration | measurand |
|---------|--------------------------|----------------|------------------|--------------------------|-----------|
|         | course and student       | set of answers | questionnaire    | mapping sets of answers to quality values | quality of course perceived by a student |
|         |                          |                |                  | synthesizing (e.g., by median or mean value) $n$ values | (median / mean) quality of course |

Case 2B directly assumes $c$ as the system under measurement, with the consequence that each $s_i$ becomes part of the measuring system together with $q$. This implies that the interaction of $q \oplus s_i$ with $c$ can now produce a value $v$ of perceived quality of $c$. In comparison with Case 2A, the main issue here comes from the much more complex structure of the measuring system, that operatively corresponds to a usual low reproducibility (“condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects ” – [8] def. 2.24): different students, and therefore different measuring systems, will produce different quality values for the same course. In a metrological perspective, this is explained in terms of an incomplete calibration of the measuring system, that has been calibrated only in its questionnaire component, but not as a whole. Hence, the problem arises of how a measuring system constituted of a questionnaire and a student who is asked to fill it can be calibrated.

The basic procedure to this goal derives from the fact that “measurement implies comparison of quantities” ([8] def. 2.1 note 2): for each question, a series of “standard” courses could be presented to
the student, each of them with an associated answer (this is structurally analogous to the standard series of Mohs’ scale), and (s)he should be instructed on how to compare (candidate) courses (against those in the series) with respect to their quality. Then, the process of filling each question could be performed as: (i) compare the course under evaluation with those in the “standard” series, and select the course in the series that is more similar to the one under evaluation from the point of view of their quality; (ii) assign to the course under evaluation the answer that is associated with the selected course in the series. An increase of reproducibility should be expected as the result of this activity: the remaining differences in the answers produced by different students could be, once again, interpreted and dealt with in metrological terms, as, e.g., measuring systems that must be recalibrated, or that must be corrected (‘correction’: “compensation for an estimated systematic effect” – [8] def. 2.53), or that are simply inadequate to the given measurement task some for reason (in case of measurement of soft quantities directly by human beings, as the present example, such reasons could include individual psychology and even ideology). On the other hand, an experimental evidence of low reproducibility could also be exploited in a feedback analysis, leading to refine the “standard” series and/or the mapping from sets of answers to quality values, and thus to partially redefine the measurand itself.

| Case 2B                  |
|--------------------------|
| **system under measurement** | **input quantity** | **measuring system** | **processing / calibration** | **measurand** |
| course                   | set of answers     | questionnaire and student | mapping sets of answers to quality values, possibly through comparison with a “standard” series | quality of course |

4. Some conclusions
The basic claim of this paper is that measurability and measurement validity are distinct issues, and as such they should be dealt with in measurement science. In the case of soft properties measurement validity is particularly critical, as the presence of human beings in the measurement process can interfere in different ways. Since questionnaires are common devices to acquire information about soft properties, they have been the object of a short case study, that has highlighted that some ambiguities can emerge in their usage as measuring instruments, which can be removed by explicitly defining an overall measurement model, comprising a measurand model and a measuring system model. Questionnaires can be used in different conditions, and in particular when the system under measurement (SuM) and the measuring system (MS) are:
(Case 1) SuM: the filling subject; MS: the questionnaire;
(Case 2A) SuM: a third entity (such as a service or a product) together with the subject; MS: the questionnaire;
(Case 2B) SuM: a third entity; MS: the questionnaire together with the subject.

While the difference between Case 1 and Cases 2 is manifest, more subtle seems the distinction between Case 2A and Case 2B. On the other hand, we have shown that such two cases can be compared with each other only if in the former the system under measurement is made independent of the subject, as typically obtained by repeating the measurement with different subjects and then synthesizing the obtained results. This shows an interesting trade-off between Case 2A and Case 2B: to get reliable information on a soft property of such entity, some resources must be devoted to perform multiple measurements or to suitably calibrate the measuring system. This latter strategy highlights that calibration can operate as a validation means, and it is worth of further exploration.
References
[1] Mari L 2003 Measurement 34 17-30
[2] Finkelstein L 2003 Measurement 34 39–48
[3] Finkelstein L 2005 Measurement 38, 267-274
[4] Rossi GB 2007 Measurement 40 545-562
[5] Krantz DH, Luce RD, Suppes P, Tversky A 1971-1990, Foundations of measurement, Academic Press
[6] Mari L 2007, Measurability, in: Measurement in Economics, Boumans M (ed.), 41-77, Elsevier
[7] Mari L, Lazzarotti V, Manzini R 2009 Measurement 42 241-253
[8] JGCM 200:2008, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), Sèvres, France, 2008; available online: http://www.bipm.org/en/publications/guides/vim.html
[9] Mari L 2010, in these proceedings
[10] JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections), Sèvres, France, 2008; available online: http://www.bipm.org/en/publications/guides/gum.html
[11] Frigerio A, Giordani A, Mari L 2009, Synthese doi:10.1007/s11229-009-9466-3