Measurement 42 (2009) 1459–1462

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

# About the treatment of systematic effects in metrology

Franco Pavese

*INRIM, strada delle Cacce 73, 1039 Torino, Italy*

ARTICLE INFO

*Article history:*
Received 27 October 2008
Received in revised form 16 July 2009
Accepted 22 July 2009
Available online 25 August 2009

*Keywords:*
Uncertainty
Systematic error
Systematic effects
Metrology

ABSTRACT

A comparison of the text of VIM recent III Edition with that of the GUM and of its contemporary VIM II Edition alights significant differences in the definition of basic measurement terms in the two documents, and with respect to the basic written standards in the field of testing, ISO 5725 and ISO 3534. The paper intends to introduce author's interpretation of these – and companion – texts, concerning specifically the terminology and the statistical treatment of the influence quantities and of the effects of their variability (in time and standard-to-standard), either related to replicated measurements performed on a single standard (standard 'reproducibility') or to the comparisons of different standards, thus involving the concept of 'accuracy' and its estimate, and consequently directly relevant to traceability.

Another question that arose a few years ago was whether different types of measurand could be the consequence of the different intrinsic nature of different types of standards. It prompted an analysis that resulted in the proposal of considering two distinct 'classes' of standards. These classes require different answers to the issue of the treatment of systematic effects. The distinction is relevant, in particular, to the statistical treatment of comparison data, which form the basis of the traceability assessment. The paper is presenting a discussion on the implications of the above distinction, concentrating on cases where systematic effects are dominating the experimental results, a common case in several metrology fields, and on ways to tackle the problem of the correction required by the GUM for standards of class 2 (standards whose values are accurate measures of a common measurand) – a class often not recognised in the general literature.



## 1. Introduction

Systematic effects are the most critical issue in measurement, especially in the metrological field, since their effect on the measured values and on the resulting total uncertainty is often dominant. In this respect, the provision for the measured values required by classical statistics, and namely by GUM [1], is that "the result of the measurement of the realized quantity is corrected for the difference between that quantity and the measurand", i.e., "it is assumed that the results of a measurement have been corrected for all recognized significant systematic effects". However, this does not generally ensure that "after the correction, the expectation or expected value of the error arising from a systematic effect is zero" because the correction cannot be made exact since "neither the value of the realized quantity nor the value of the measurand can ever be known exactly"; or, because the correction is, at best, applied only to "all recognized significant systematic effects", while the presence of non-recognised systematic effects cannot generally be excluded [1,2]. In fact, it is common experience that (*inter*-laboratory) standard comparisons are necessary operations to perform as the only possible way to detect otherwise unrecognised

*E-mail address:* f.pavese@inrim.it

0263-2241/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.measurement.2009.07.017

significant differences between the values of standards or of tests; and, it is a shared experience the fact that even for well-behaved experiments those comparisons *do show* the existence of such significant differences (e.g., more than 50% of the key comparisons presently available on the BIPM KCDB [3]).

One consequence is that in the majority of cases a zero-mean distribution underlying the data, assumed to be one of the basis for the GUM approach, is not achieved and that equally not achieved is the assumption of "repeated measurements", namely in the case of the treatment of *inter*-laboratory comparison data.

If "repeatability conditions" cannot be ensured to hold,[1], most commonly available statistics for the treatment of these data do not hold, being based on repeated measurement conditions. This is largely recognised and the solution indicated is invariable that one should remove the systematic effects before data can be statistically treated. This approach starts since Gauss [4,5] "explicitly excluding the consideration of systematic [*regular, in his terminology*] errors in his investigation and warns that 'of course, it is up to the observer to ferret out all sources of constant error and remove them' ". See similarly in the recent [6] "We saw that within the framework of conventional statistics, the measured quantity is conveniently written as $X = v_t + M$, where $v_t$ is the unknown true value of $X$ and $M$ is the measurement error. The formalism [*in conventional statistics*] applies if $E(M) = 0$. But this condition holds only in the case when no systematic effects are assumed to influence the 'errors' in acquiring data $x$ under repeatability conditions. A systematic effect occurring during the series of measurements is neither random nor observable. It introduces a bias that may be estimated from other information, mostly in the form of enclosing limits, cannot be exactly known. Thus, it constitutes an element for which treatment conventional statistics fails utterly".

However, systematic effects do occur and the experience shows that it is a wishful thinking to believe to always be able to remove all of them (one of the argument for 'metrology' to be an 'art' based on 'expert judgment').

Unfortunately, a firmly based theory for their treatment is not presently available, and not even there is consensus on the intrinsic nature of systematic effects,[2] namely their being of random nature or not. In the literature there is a vast range of definitions for both viewpoints (see [7] for a selection of them). Looking at the two basic Guides in metrology, GUM (1993) and VIM (2008) [8], their approaches are different. For GUM, after correction they are zero-mean random variables, whose contribution to uncertainty is to be added to the 'random' component.[3],[4] Instead, VIM is defining "systematic measurement error" as "component of measurement error that in replicate measurements remains constant or varies in a predictable manner", so not being a random variable.[5]

The above considerations have obvious consequences on the meaning and use of the concept of 'accuracy'[6] and consequently on the concept of "metrological traceability". On the other hand, one should note that the latter concept is of a hierarchical nature. Instead, when talking of comparisons of standards between NMIs, one can only talk of "degree of equivalence", a non-hierarchical concept whose need arises from the lack of availability of any further upper level in the traceability chain. In fact, the key comparisons are *inter pares* exercises [9]. In the rest of the paper we will therefore talk instead of "comparability".[7]

## 2. Does comparability mean the same for all types of standards?

Obviously, the first requirement for comparability to exist is that the measurand of the comparison is clearly defined and intended to be the same for all measurements providing the data to be compared. This may be not so trivial, as discussed in [11] and the problem is well known in chemical metrology and for the bio-medical quantities.

In all instances, an analysis of the different types of standards, has brought in the past years to the need of making some basic distinctions based on their intrinsic nature and to propose two basic class of them [10,12–

[1] Actually, by saying that are valid "*over a short period of time*" the definition of repeatability is a tautology, since in fact it means 'short enough to ensure that repeatability conditions hold".

[2] It is worth noting that the influence quantities are all and the only potential causes of systematic errors (incidentally, almost never time is an influence quantity in itself, but in time the influence quantities can show a variability).

[3] This is not explicitly indicated in GUM, where the uncertainty is not classified in terms of 'random' and 'systematic' components, but as Type A (determined by statistics) and Type B (determined by other means) components and where no reference is made to the terms 'reproducibility' and 'accuracy'. This leaves the user in some confusion. In fact, GUM about Type A is also stating "Variations in repeated observations are assumed to arise because influence quantities that can affect the measurement result are not held completely constant", a definition that universally applies instead to reproducibility: the concept of "influence quantity" *variability* is not a characteristic of 'repeatability' ("repeated observations") but of 'reproducibility'. On the other hand, the later VIM (2008) is defining Type A evaluation of measurement uncertainty as arising from "evaluation of a component of measurement uncertainty by a statistical analysis of quantity values obtained under defined measurement conditions", where the conditions can be "*repeatability* condition of measurement, *intermediate precision* condition of measurement and *reproducibility* condition of measurement" (for their statistically-based part).

[4] But the model incompleteness, considered by GUM as a random component in the model, should instead be considered as a 'non-uniqueness' of the definition, not of random nature nor producing random effects.

[5] But with some contradictions. E.g., in the Note 1 to this definition, it is indicated "A reference quantity value for a systematic measurement error is a *true quantity value*, or a *measured quantity value* of a measurement standard of negligible measurement uncertainty, or a *conventional quantity value*": however, the first value is unknowable, the second value is uncertain, the third value is not comparable directly with the measured value and may be uncertain.

[6] Together with the discussion about "unknowable quantities" such as the "true value", not treated in this paper: in this respect see [10]. The present paper will not treat either the concept of 'trueness'.

[7] Notice that the term "metrological comparability" will not be used in this paper since it has been given by VIM (2008) the different meaning of "comparability of measurement results, for quantities of a given kind, that are metrologically *traceable* to the same reference": apparently VIM does not take the special issue of the NMIs into consideration.