International Conference on Metrology of Environmental, Food and Nutritional Measurements
three scientific events at the same place and time
2nd IMEKO TC19 Conference on Environmental Measurements
1st IMEKO TC23 Conference on Food and Nutritional Measurements
1st IMEKO TC24 Conference on Chemical Measurements

10 – 12 September 2008, Budapest, Hungary,

A METROLOGIST ANALYSIS OF INTERNATIONAL WRITTEN STANDARDS AND GUIDES RELEVANT TO MEASUREMENT IN CHEMISTRY

Franco Pavese [1]

[1] INRIM, Torino, Italy, f.pavese@inrim.it

Abstract: This paper, starting from reviewing basic concepts, such as repeatability, reproducibility, accuracy, systematic error, true value, bias, Type A and B uncertainty components, as defined in international documents, shows that, currently, full consensus and a common understanding seems not be achieved yet. This fact reflects on the process of combining data, namely in chemistry.

Keywords: repeatability, reproducibility, accuracy, Type A uncertainty component, Type B uncertainty component.

1. INTRODUCTION

The degree of consistency of different measurement results, obtained by different independent experimenters or by the same experimenter at different times, is considered in general to provide a measure of the degree of reliability that can be associated to the results in representing the quantity under study, having taken into account the fact that experimental knowledge is always imperfect to some degree.

Consequently, replication of measurements and combination of observations are standard and essential practices in science.

This requirement places the need to evaluate the degree to which different observations can correctly or safely be compared (or combined) with each other, or, in other words, to assess the *traceability* of the measurements performed by different experimenters and at different times.

An underpinning basic concept of science, and hence of measurement science, is that, due to imperfect knowledge of the observed phenomena, the numerical data that are the outcomes of measurement are affected by errors. Irrespective to the reasons that are the causes of these errors, the resulting dispersion of the numerical values that is generally observed is interpreted as an evidence of the imperfect knowledge.

Thus, the dispersion of the measured data introduces an uncertainty in the measure of the observed phenomena. Uncertainty associated with data is specified according to models that are different according to the underpinning assumptions, which must adequately match the characteristic of the observed phenomena or process.

2. INTERNATIONALLY RECOGNISED APPROACHES ON THE EXPRESSION OF UNCERTAINTY IN MEASUREMENT

In evaluating the uncertainty associated with measurement results in metrology and testing, several steps can be enumerated, each characterised by the use of different methods to fullfil correspondingly different purposes [1, 2].

For a measurement process entirely taking place within a single laboratory, the purpose of measurement replication,

(a) when performed on the same measurement standard, is primarily to obtain statistical information providing a measure of the *repeatability* of the measured values of the standard;

(b) when performed on the same measurement standard, is then to evaluate the increase in the total uncertainty arising from the variability of the influence quantities affecting the standard, including those that have a dependence on time, i.e. to have a measure of the *reproducibility* of the measured values of the standard;

(c) when performed on several measurement standards of the laboratory, is finally to assess whether they have the same value or to provide a measure of the (systematic) differences between their measured values, and to evaluate the associated uncertainty, i.e. to provide an estimate of the *accuracy* of the measured values of the laboratory standards. This step is called *intra*–laboratory comparison.

When operation (c) is performed by directly comparing one (or more) measurement standards provided by different laboratories, so that it is part of a process taking place between *at least two* laboratories, it is then called an *inter*–laboratory comparison.

When the same operation is performed to assess "periodically the overall performance of a laboratory" [3], i.e. to show that the laboratory can continue to demonstrate its ability to conduct correctly a certain type of measurement, it should be considered and used as a *proficiency test* (PT).

Most of the written standards, concerning not only testing by also metrology, are therefore defining the terms corresponding to the three basic steps of any traceable measurement process (repeatability, reproducibility and accuracy) and are instructing about performing them.

However, since the 1980's a different process started and took place, ending up in 1995 with the publication by ISO of a joint *Guide for the expression of uncertainty in measurements* (GUM) [4], that does not make use of these terms for the random and systematic components of uncertainty underpinning the concepts of repeatability, reproducibility and accuracy. It resorts instead on the distinction of the uncertainty components in Type A and Type B, depending on the use or not of statistical methods for their evaluation.

The ambition was to replace in metrology and in testing the other approaches, having been established by a group of experts pertaining to a wide range of International Organisations. After more than 10 years it is hard to say if it fully succeeded, especially in fields like chemistry and biology that have to handle very specific problems. The paper is first illustrating author's opinion about this issue based on the nomenclature and guidelines found in International written standards and Guides published –and often updated– from the time of the issue of GUM on.

3. DEGREE OF CONSISTENCY OF THE WRITTEN STANDARDS AND GUIDES PRESENTLY IN FORCE

In order to give here a flavour of the problems that a previous author's analysis has found [5], the following statements are reported, which does not indisputably seem to always be in accordance with each other.

3.1 Repeatability

According to ISO 3534-2 (2006) [6], "repeatability conditions" of measurement are "observation conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in the same test or measurement facility by the same operator using the same equipment within short intervals of time".

The VIM (2008) [7] definition (2.11) adds to "condition of measurement in a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and location ... over a short period of time " also "and replicate measurements on the same or similar objects".

3.2 Reproducibility

According to the VIM (2008) (2.24) a reproducibility condition is a "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". Further, it notes, "the different measuring systems may use different measurement procedures".

In testing, ISO 5725 [8] (3.18), which also includes the effects of calibration and time, does not allow for different measurement procedures be used.

For QUAM [9] (2000) reproducibility is instead the "variability obtained when different laboratories analyse the same sample" and it uses the above definition for the intermediate precision.

3.3 Accuracy

The concept of accuracy involves that of *systematic effects*. However, the meaning of the latter is not inequivocally specified in recent written standards and in the literature.

In the VIM (2008) (2.18), a systematic effect is a "component of measurement error that in replicate measurements remains constant or varies in a predictable way", changing a definition that in 2004 was still the "mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions minus a true value of the measurand" (i.e., a random variable).

Similarly, in QUAM the systematic error is defined as "a component of error which, in the course of a number of analyses of the same measurand, remains constant or varies in a predictable way. *It is independent of the number of the measurements made and cannot therefore be reduced by increasing the number of analyses under constant measurement conditions*". ISO 21749 [10] states "sometimes it is difficult to distinguish a systematic effect from random effects and it becomes a question of interpretation and the use of the related statistical models. In general, it is not possible to separate fully random and systematic effects".

In ISO 5725 and, in general, in testing documents the term *bias* is most commonly used, in fact, instead of systematic error. ISO 3534-2 (2006) states: (1.3.3.2) "bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the true value is reflected by a larger bias value". *Bias* is also defined as follows: "the difference between the expectation of the test results and an accepted reference value"; "the bias of a test method is usually determined by studying relevant reference materials or test samples . . . The uncertainty associated with the measurement of the bias is an important component of the overall uncertainty" (EA 4/16) [3]; "where the bias is significant compared to the combined uncertainty, additional action is required" (i.e., eliminate, correct, report or increase uncertainty) (QUAM).

Furthermore, DIN 1319–1 [11] indicates that the total systematic error comprises two components:

— one covers the *known* systematic measurement error,

— the other the *unknown* systematic measurement error.

The GUM [4], as illustrated here below, clearly considers only the *known* systematic errors, i.e., the recognised effects of some influence parameters, with an approach consisting in randomising the recognised systematic effects.

Coming now to the term *accuracy*, the VIM (2008) (2.13) states: "closeness of agreement between a measured quantity value and a true quantity value of a measurand".

The *true value* of a quantity in the VIM (2008) (2.11) is the "quantity value consistent with the definition of a quantity", noting that "in the Error Approach to describing

measurement, a true quantity value is considered unique and, in practice, unknowable. The Uncertainty Approach is to recognize that, owing to the inherently incomplete amount of detail in the definition of a quantity, there is not a single true quantity value but rather a set of true quantity values consistent with the definition. However, this set of values is, in principle and in practice, unknowable".

Testing applications very often show an intrinsic difference with respect to most metrology applications, in that a *true value* can be assigned to the measurand: in fact, ISO 5725 defines *trueness* (3.7) in an operational way as "the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value".

3.4 Type A component of uncertainty

According to the GUM [4], Type A is a "... method of evaluation of a standard uncertainty by the statistical analysis of a series of observations" (2.3.2). The GUM, apparently drops the term 'repeated' from the definition with respect to INC–1 [13], but does in fact use and refer to 'repeated', e.g., in (3.3.5): ". . . obtained from a Type A evaluation is calculated from a series of repeated observations ...".

However, the GUM in (3.1.5) specifies, "variations in repeated observations are assumed to arise from not being able to hold completely constant each influence quantity that can affect the measurement results", the same in (3.2.2).

In addition, the GUM in (3.2.4) prescribes: "it is assumed that the results of a measurement have been corrected for all recognised significant systematic effects", effects that are arising from "not being able to hold completely constant each influence quantity".

The GUM prescription is typical of the replication of measurements for the purpose (b), i.e., to get a measure of the *reproducibility*.

For an *influence quantity*, the GUM means a "quantity that is not included in the specification of the measurand but that nonetheless affects the result of the measurement" (2.7), while in the VIM (2008) it is defined as a "quantity that, in a direct measurement, does not affect the quantity that is actually measured, but affects the relation between the indication and the measurement result" (2.52).

According to the VIM (2008), "Type A evaluation of measurement uncertainty" arises from the "evaluation of a component of measurement uncertainty by a statistical analysis of measured quantity values obtained under defined measurement conditions" (2.28), where the conditions can be "repeatability condition of measurement, intermediate precision condition of measurement, and reproducibility condition of measurement", so including also intermediate and reproducibility conditions.

3.5 Type B component of uncertainty

According to the GUM, Type B is a "... method of evaluation of a standard uncertainty by means other than the statistical analysis of a series of observations" (2.3.3). "Type B evaluation" ("of measurement uncertainty components") is "founded on a priori [probability] distributions".

The VIM (2008) (2.29) definition of Type B is: "evaluation of a component of measurement uncertainty determined by means other than a Type A evaluation of measurement uncertainty", with examples given: "evaluation based on information associated with authoritative published quantity values; associated with the quantity value of a certified reference material; obtained from a calibration certificate; about drift; obtained from the accuracy class of a verified measuring instrument; obtained from limits deduced through personal experience". These examples of Type B evaluation seem basically involving only expert judgment.

3.6 Overall GUM approach

Neither Type A nor Type B evaluations, as defined by the GUM, seem to fit unequivocally measurements performed for purpose (b), the assessment of *reproducibility*.

In the GUM (B.2.3) "the term 'true value' is not used" since it is "viewed as equivalent" to the term "value of a measurand", and it cannot be determined. However, for the definition of accuracy, the VIM 2st edition one is adopted, resorting to the term *true value* (this term is no more used by the VIM 3rd edition (2008), in order to ensure the VIM consistency with both GUM and IEC 60359 [14,15]

EA–4/16 [3] indicates "this document interprets the GUM as based on corrections included in the model to account for systematic effects; such corrections are essential to achieve traceability".

A2LA Guide [16] uses the term *bias*, adding: "the [GUM] method assumes that all significant systematic effects have been identified and either eliminated or else compensated for by allocation of suitable corrections".

4. APPLICATION OF THE WRITTEN STANDARDS AND GUIDES AND DATA MODELS

On the light of the above document citations, there are effects on the procedures for combining data and ensuring *traceability of assigned values*, namely in chemistry.

First of all, there is the problem of some observed inconsistencies, which arose in the evolution of some of the concepts in the last 15 years, mainly due to different approaches having being selected for different documents.

The GUM is not using the term 'true value', though it is using the concept, for illustrating, e.g., how the "result of measurement" and the "final result of measurement" are obtained. IEC 60359, on the contrary is stating that the very concept of 'true value' is unnecessary for its goal in measurement that is to obtain consistent results. In ISO testing documents, instead, it is normally underlying the concept of 'reference value' or of 'consensus value', so not being consistent with the VIM 3rd edn. definition.

There are major differences in the concept of "systematic effect" or "systematic error" (the term 'error' is not used in GUM). For the GUM, it is a random variable and a correction *must* be applied for all recognised significant systematic effects (for QUAM [9] also for not significant

ones), so that after correction, the expectation of the systematic effects is zero. The correction requirement is also common in testing. However, in metrology, being the 'true value' normally unknowable[1], and being any experimental knowledge affected by uncertainty, an exact correction is not possible. Consequently, the "zero expectation after correction" condition is not reachable. One consequence is that in the GUM data modelling does not require a 'bias' term be included, as we will illustrate later on, since 'bias' is identically equal to zero.

In the VIM (2008), the systematic error is said to remain constant or change in a predictable way: in general, this would be a condition only holding *over a short interval of time* and only under *repeatability* conditions, while under *reproducibility* conditions is supposed to vary by definition. It is therefore unclear how to reconciliate this definition with the others, where effects variability occurs.

Further, the GUM is entirely based on *repeated* measurements (in addition to non-statistical Type B information), but its definition of *repeatability* is contrasting any other definition that can be found in International documents, when it says, "it is assumed that the results of a measurement have been corrected for all recognised significant systematic effects". This is normally the definition of *reproducibility* conditions, except in QUAM (see above).

Coming now to the data models that are derived from the different approaches illustrated so far, there is a basic difference between the GUM and the other documents, namely those intended specifically for use in testing.

The GUM data modelling –not to be confused with the GUM model described in (4.1) for an indirect measurement– is based on the fact that only *repeated* measurements are treated by the GUM, with no distinction, "after correction", between the random errors and the (variability of) the systematic errors. It can be written, for the i-th of total I observations as:

$$y_i = a + \varepsilon_i \qquad i = 1, \ldots, I \tag{1}$$

where y_i is an observation drawn from a random variable Y, a is the realised quantity value and ε_i is the total random error –including both random error and the zero-mean variability of the systematic effects. Notice that any possible bias is not modelled because it is assumed to be identically equal to zero. Should the assumption of repeated measurements not be found true by a one-way consistency test[2], model (1) would simply not apply, with no provision about what action should be undertaken, except checking the reason for the inconsistency and correcting it –not always allowed, e.g. in MRA [12] key comparisons.

When on the contrary the condition of zero bias is not assumed to occur, but, instead, the existence of bias is assumed as a *prior* knowledge, consisting of the evidence that, namely for comparisons and PTs in general, ''when the i-th participant repeats the comparison j times, then its results can be distributed about an expectation value differing from the measurand value a by an amount b_i with standard deviation s_i'' [18], b_i has the meaning indicated in the following model (2), for the i-th of total I observations in the j-th of total J laboratories, basically the one prescribed in ISO 5725:

$$y_{ij} = a + b_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, I \quad j = 1, \ldots, J \tag{2}$$

where "a is the general mean (expectation); b is the value of the laboratory component of bias under repeatability conditions; ε is the random error occurring under repeatability conditions" [8]. Model 2 can also be written by adding the "bias of the measurement method," d, to b_i [8] or by explicitly specifying several contribution $b_{ij} = \sum_j b_{ij}$. Commonly in testing, a is known, assessed by a hierarchically higher rank of laboratories ('reference value') or stipulated by consensus. In this case, 'bias' is modelled – 'bias' is a term that may be improper to use in metrology, though convenient because short instead of the long, but correct, sentence in [18]. Then, after estimates of the differences between values b_i are obtained[3], a check for the compatibility of the b_i with each other may be performed. Compatibility-test failure for some of the $(b_h - b_k)$ [4] shall indicate that the hypothesis that these values are not significantly different from zero is false.

In chemical measurements, model (2) is largely preferred in testing –where it is the only adopted by all the different international documents and Guides. This is due to the fact that, normally, a 'reference value' or 'consensus value' is available. In this frame, the difference between a 'consensus value' and a 'reference value' is that, the latter is an information *a priori* derived from characterisation measurements of a batch of material, the former is directly obtained with a *consensus statistical procedure* from the participants' results [19]. Therefore, the bias, for which a correction is then usually applied, can precisely be known.

[1] There are few exceptions. E.g., when the use of an 'ideally pure substance' is specified, the true condition of impurity concentration $x = 0$ is known.

[2] Usually, a χ^2-test is proposed for this purpose or the use of ''normalised errors'' (''metrological ratio'') or z-score (e.g. see [17] and references therein). Test failure shall not pass the hypothesis that the measurements are repeated. Test acceptance shall not change the intrinsic non-repeated nature of the measurements involved. Consequently, this method is generally likely to underestimate the uncertainty associated to the measured differences.

[3] In fact, in metrology, the b_i remain as unknown as a is, only the differences $(b_h - b_k)$ of pairs of laboratories are measured.

[4] According to the VIM [7], the definition of ''metrological compatibility'' is (2.47) as the ''absolute value of the difference of any pair of measured quantity values from two different measurement results is smaller than some chosen multiple of the standard measurement uncertainty of that difference,'' also noting that the ''metrological compatibility of measurement results replaces the traditional concept of ''staying within the error,'' as it represents the criterion for deciding whether two measurement results refer to the same measurand or not. If in a set of measurements of a measurand, thought to be constant, a measurement result is not compatible with the others, either the measurement was not correct (e.g. its measurement uncertainty was assessed as being too small) or the measured quantity changed between measurements.''

On the contrary, in metrology a difference between the values of different standards or different laboratories can be detected only after a comparison is performed. However, the metrology rules in the frame of the MRA, while providing a "degree of equivalence" between the laboratories (a non-hierarchical concept, so not corresponding to that of *traceability* that is instead a hierarchical one), is not allowing any use of this equivalence, except when "systematic unresolved differences" (SUD) remain, in which case they should be resolved by a corresponding increase of the stated uncertainty of the relevant participants. However, this issue is still under scrutiny in most practical cases and was almost never taken in consideration, though a recent statistics of the more than 200 key comparisons (KC) today published by BIPM [20], have shown that for more than half of the KCs the results do not pass the usual tests of consistency concerning the suitability of model (1).

Some final considerations are useful about the use of Type B components of uncertainty, as defined by the GUM, i.e., using information other than the statistical one.

Type B is often understood as to deal with systematic error, while Type A is assumed to deal with only random error, or that only Type B is dealing with *prior* knowledge. Neither of these assumptions is true. As indicated before, Type A also includes the random component of uncertainty arising from the systematic effect variability after correction. On the other hand, there is *prior* knowledge that should be included in Type A. For example, for a standard having proved to be stable in time, one can increase by this way the number of the *repeated* observations available for that specific standard over the years, which can then be statistically analysed all together, and the uncertainty be considered as part of Type A components. Therefore, Type A components of uncertainty can also embed *prior* information. Incidentally, this can be handled using statistics, so that it is not true that Bayesian methods are the only effective in taking *prior* knowledge into account.

Prior knowledge can also include many other different types of *prior* knowledge. A short, non-exhaustive, list can be found in the VIM (2008):

a) "associated with authoritative published quantity values;
b) associated with the quantity value of a certified reference material;
c) obtained from a calibration certificate;
d) about drift;
e) obtained from the accuracy class of a verified measuring instrument;
f) obtained from limits deduced through personal experience."

Thus, *prior* information is considered in (c) the calibration certificate of a device, in (b) the reference value of a batch of material, or in stating the degree of equivalence originated by a MRA exercise [12]. To these values also an uncertainty is associated and possibly, in addition, a probability distribution (empirical, if assigned on the basis of the results of the original studies), most often Gaussian. For the user, they act as being assigned (or stipulated), so loosing their possible original content of subjectivity that could have contributed to their determination when the original studies were performed.

Similarly one should consider item (a) and (e). On the contrary, item (f) should be considered as an example of an 'expert judgement'. Also in many of those cases, Bayesian methods are not the only effective way for taking *prior* knowledge into account, e.g., when the uncertainty is expressed as bounded in a confining interval [21].

REFERENCES

[1] F. Pavese and E. Filipe, "Some metrological considerations about replicated measurements on standards", Metrologia 43, 2006, 419–425

[2] F. Pavese, "An Introduction to Data Modeling Principles in Metrology and Testing", Ch.1 in "Data Modeling for Metrology and Testing in Measurement Science", Series "Modeling and Simulation in Science, Engineering and Technology", Birkhauser, Boston, pp 1–30 of pp 501, ISBN: 978-0-8176-4592-2, to appear on October 2008

[3] European Accreditation, "EA guidelines on the expression of uncertainty in quantitative testing", EA-4/16, December 2003, rev00

[4] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, "International Vocabulary of Metrology" (GUM), 2nd edn. International Organization for Standardization, Geneva, Switzerland, 1995

[5] F. Pavese, "Replicated observations in metrology and testing: modelling repeated and non-repeated measurements", Accred. Qual. Assur. 12, 2007, 525–534

[6] ISO 3534-2 (1993) 2nd edn and (2006) 3rd edn "Statistics—vocabulary and symbols —Part 2: applied statistics", International Organization for Standardization, Geneva, Switzerland

[7] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, "International vocabulary of basic and general terms in metrology", (VIM) 3rd edn, 2008

[8] ISO 5725 "Accuracy (trueness and precision) of measurement methods and results", International Organization for Standardization, Geneva, Switzerland, 1994

[9] Eurachem CITAC, "Guide CG4. Quantifying uncertainty in analytical measurements" (QUAM 2000.1), 2nd edn, 2000

[10] ISO 21749 "Measurement uncertainty for metrological applications—simple replication and nested experiments", International Organization for Standardization, Geneva, Switzerland, 2003

[11] DIN 1319-1 "Fundamentals of metrology—Part I: basic terminology", Berlin, Beuth,1995

[12] CIPM, "Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes", Bureau International des Poids et Mesures, Sèvres, 1999

[13] R. Kaarls, Proc.Verb. Com. Int. Poids et Mesures 49: A1–A12 (French), 1981; P. Giacomo P Metrologia 18: 43–44 (English), 1981

[14] CEI/IEC 60359 "Electrical and electronic measurement equipment –Expression of performance" 2001

[15] C. Ehrlich, R. Dybkaer and W. Wöger, "Evolution of philosophy and description of measurement (preliminary rationale for VIM3)" Accred. Qual. Assur. 12, 2007, 201–218

[16] American Association for Laboratory Accreditation (A2LA) "Guide for the estimation of measurement uncertainty in testing", 2002

[17] A.G. Steele and R.J. Douglas, "Simplicity with advanced

mathematical tools for metrology and testing", Measurement 39, 2006, 795–807

[18] D.R. White, CPEM, Sydney, Australia, CPEM Conference digest, pp 325–326, 2000

[19] A. Baldan, A.M.H. van der Veen, D. Prauß, A. Recknagel, N. Boley, S. Evans and D. Woods, "Economy of proficiency testing: reference value versus consensus value", Accred Qual Assur 6, 2001, 164-167

[20] See http://kcdb.bipm.org

[21] V. Kreinovich, "Interval Computations and Interval-Related Statistical Techniques: Tools for Estimating Uncertainty of the Results of Data Processing and Indirect Measurements", Ch.4 in "Data Modeling for Metrology and Testing in Measurement Science", Series "Modeling and Simulation in Science, Engineering and Technology", Birkhauser, Boston, pp 119–148 of pp 501, ISBN: 978-0-8176-4592-2, to appear on October 2008. Additional information in the attached DVD.