Accred Qual Assur (1998) 3:180–184

GENERAL PAPER

Jean Pauwels
Andrée Lamberty
Heinz Schimmel

The determination of the uncertainty of reference materials certified by laboratory intercomparison

Received: 30 September 1997
Accepted: 7 December 1997

Jean Pauwels (✉) · Andrée Lamberty
Heinz Schimmel
European Commission,
Joint Research Centre
Institute for Reference Materials and Measurements (IRMM)
B-2440 Geel, Belgium
Tel.: +32–14–571722
Fax: +32–14–590406
e-mail: pauwels@irmm.jrc.be

Abstract A pragmatic method is proposed for the implementation of the Guide to the expression of uncertainty in measurement in the certification of reference materials by laboratory intercomparison. It is based on the establishment of a full uncertainty budget for each laboratory result and the estimation of the impact of various laboratory standard uncertainties and of between-units variability on the certified reference material (CRM) uncertainty.

Key words Reference material · Laboratory intercomparison · Certified value · Uncertainty

Introduction

Many reference materials, produced worldwide, are certified by laboratory intercomparison, involving a large number of independent and, if possible, equally competent laboratories [1]. Normally, methods used are based on a variety of chemical and/or physical principles. It is then assumed that the differences between individual results, both within and between laboratories, are all of a statistical nature regardless of their causes. Each laboratory mean is considered as an unbiased estimate of the property of the material to be certified, and usually an unweighted mean of the laboratory means is assumed to be the best estimate of that property. In general, a reference material certification involves different laboratories, each of which measures the requisite property on different samples, with each sample measurement consisting of a number of independent repeated observations. The certified value and its uncertainty are then estimated on the basis of an analysis of variance, after verification that all data belong to the same normally distributed population.

If this is the case, the mean value of all individual data is taken as the certified value, and the half-width of the 95% confidence interval of the mean value of all individual data as its uncertainty. If, on the contrary, pooling is not allowed because individual data do not belong to the same normally distributed population, the mean value of the laboratory means is taken as the certified value and the half-width of the 95% confidence interval of the mean value of the laboratory means as its uncertainty.

The limitation of such procedures is that the distribution of the considered values should be normal and that no other sources of uncertainty than "random experimental uncertainties" should exist [1].

The above procedure finds its justification in the fact that one presumes that, if a large variety of independent laboratories and methods is used, possible systematic effects in the individual laboratory results will be "randomized" and that, eventually, both the residual systematic error and its uncertainty are reduced to zero.

Determination of an uncertainty according to the Guide to the expression of uncertainty in measurement

According to the Guide to the expression of uncertainty in measurement (GUM)[2], the result of a measurement corresponds to the estimate of the value of a measurand and should, therefore, always be accompanied by an uncertainty statement. It is, generally, deter-

mined on the basis of a series of observations obtained under repeatability conditions; its standard uncertainty is expressed as a standard deviation. It is assumed that measurement results are corrected for recognized significant systematic effects and that every effort has been made to identify and quantify such effects. Moreover, any other sources of uncertainty should be estimated and taken into account.

Uncertainty components are of two different types based on the method used for their evaluation: *type A* uncertainties are evaluated statistically on the basis of a series of observations, and *type B* uncertainties on the basis of all means other than statistical ones (e.g. previous experimental data, knowledge or experience, manufacturer's specifications, data from certificates, published reference data, etc). Both *type A* and *type B* uncertainties can be of a *"random"* as well as of a *"systematic"* nature.

A measurand Y is, however, generally not measured directly, but determined from N other quantities X_1, X_2, ..., X_N through a functional relationship f:

$$Y = f(X_1, X_2, ..., X_N) \quad (1)$$

The set of input quantities X_1, X_2, ..., X_N may be categorized as

– quantities whose values and uncertainties are directly determined in the current measurement; they may then be obtained from a single observation, repeated observations or judgement based on experience; they may involve the determination of corrections to instrument readings and corrections for influence quantities

– quantities whose values and uncertainties are brought into the measurement from external sources, such as quantities associated with calibrated measurement standards, certified reference materials, reference data obtained from handbooks, etc.

The estimated standard deviation associated with the output estimate y of Y, termed *combined standard uncertainty* and denoted $u_c(y)$ is determined from the estimated standard deviation associated with each input estimate x_i of X_i, termed *standard uncertainty* and denoted $u(x_i)$. In its second recommendation, the Comité International des Poids et Mesures (CIPM) requested that this combined standard uncertainty be used "by all participants in giving results of all international comparisons or other work done under the auspices of the CIPM and Comités Consultatifs" [3].

Although $u_c(y)$ can be universally used to express the uncertainty of the result of a measurement, it may be required to give a measure of uncertainty that defines an interval about the measurement result that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand. This additional measure is termed the *expanded uncertainty* and is denoted U. It is obtained by multiplying $u_c(y)$ by a *coverage factor k*:

$$U = k \cdot u_c(y) \quad (2)$$

The generally chosen value of the coverage factor k is 2 or 3. If the probability distribution characterized by y and $u_c(y)$ is approximately normal and the effective degrees of freedom of $u_c(y)$ of significant size, $k = 2$ or 3 corresponds to a level of confidence of approximately 95 or 99%.

The result of a measurement is conveniently expressed as:

$$Y = y \pm U \quad (3)$$

which means that the best estimate of the value attributable to the measurand Y is y, and that

$$y - U < Y < y + U \quad (4)$$

is the interval that may be expected to encompass a large fraction (p) of the distribution of values that could reasonably be attributed to Y. The fraction p of the probability distribution is named *coverage probability* or *level of confidence*.

The Eurachem document "Quantifying Uncertainty in Analytical Measurement"[4] shows how the GUM concept should be applied in chemical measurement and illustrates this by four worked examples. These examples are however limited to simple analytical determinations, and the document discusses neither the problem of laboratory intercomparisons nor their use for the certification of reference materials.

Application of the GUM to the determination of the uncertainty of CRMs by laboratory intercomparison

A typical example of a certification exercise by laboratory intercomparison (e.g. for BCR CRMs) is shown in Fig. 1:

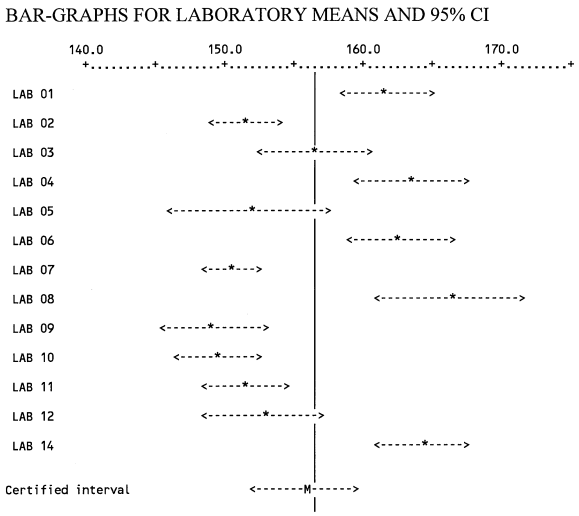

Fig. 1 Example of certification by laboratory intercomparison as performed to-day

– Between 6 and 15 laboratories carry out each six measurements spread on two different units.
– Samples of each of both units are measured on two different days.
– The measurement is (e.g. for BCR CRMs) carried out *under reproducibility conditions*, i.e. such that each replicate has its own calibration, dissolution, extraction, blank determination, etc.

The comparison of the results is, however, limited to the bare values of the six replicates carried out by each laboratory, with the immediate consequence that laboratories very often do not overlap between each other. Frequently, it is observed that the results of several laboratories participating in the certification do not even overlap with the value which is certified. The reason for this is not, as is generally believed on the basis of routine statistical tests, that there are significant differences between the results of the different laboratories, but because only the standard uncertainty on the six replicates is considered and because calculation of a combined standard uncertainty for each participating laboratory result is omitted. As already indicated, each analyst carrying out a measurement should always make up a complete uncertainty budget considering all *recognized* components of standard uncertainty affecting his measurement result. This should a fortiori apply to any laboratory which is invited to contribute to the certification of a reference material. The standard deviation $s(j)$ of the six replicates carried out by laboratory j, further denoted as $u_1(j)$ already includes part of the uncertainties of a purely statistical nature due to day-to-day variation, calibration (at least if each replicate has its own calibration), recovery yield (same remark), etc, as the measurements are in principle executed under *reproducibility conditions*. However, the standard uncertainties $u_i(j)$ (for i ranging from *2* to n) due to sampling, dry mass determination, calibration, recovery yield, blank correction, matrix effect, possible interferences, etc, generally also contain components of a more systematic nature which are not included in $s(j)$ and which are in general of a much larger magnitude. These should then as well be taken into account in the calculation of the combined uncertainty $u_c(j)$ and the expanded uncertainty $U(j)$ of each laboratory result:

$$U(j) = k \cdot u_c(j) = k \cdot \sqrt{\sum_{i=1}^{n} [u_i(j)]^2} \qquad (5)$$

i = identification number of all uncertainties considered in each individual laboratory j, varying from 1 to n, with n not necessarily identical for each laboratory

From this moment on, it can be assumed that all laboratory results are corrected for recognized significant systematic effects, that every effort has been made to identify and quantify them, and that all sources of uncertainty have been estimated and taken into account.

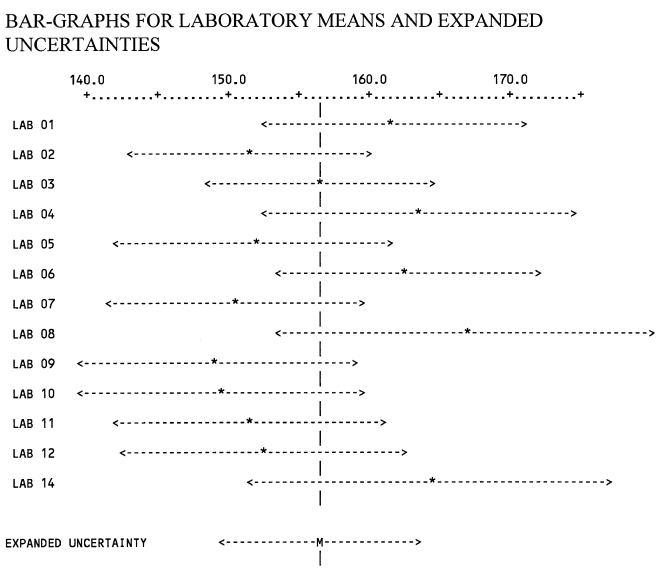

Fig. 2 Example of certification by laboratory intercomparison with consideration of combined standard uncertainties

Therefore, all results should in principle overlap and any discrepancies as shown in Fig. 1 should no longer exist (see Fig. 2). At this point, it should however be noted that components of standard uncertainty which are not laboratory specific but which are common to all or to part of the participating laboratories (e.g. those using identical methods) should be considered separately. For this reason it is essential that each laboratory supplies the project leader with a fully detailed uncertainty budget and that these uncertainty budgets are extensively discussed with the experts of all participating laboratories.

The certified value can then be calculated as either the unweighted or as the weighted mean of the laboratory means. In principle the former should be preferred, but in practice it may be unfair towards some laboratories, as especially type B components of uncertainty may have been evaluated differently from one laboratory to another. The certified uncertainty can be calculated after deconvolution (and later recombination) of all laboratory standard uncertainties in distinct categories of (combined) standard uncertainties, which may be evaluated as type A and/or as type B:

1. uncertainties which are *exclusively laboratory-dependent* $[u_c(I)]$

These affect the certified uncertainty interval in such a way that the more laboratories are involved in the intercomparison the smaller their contribution becomes:

$$u_c(I) = \frac{\sqrt{\sum_{j=1}^{l} [u_c(j)]^2}}{l} \qquad (6)$$

j = laboratory identification number, varying from 1 to l
l = total number of laboratories

2. Uncertainties which are *common to all laboratories* participating in the certification $[u_c(II)]$
These affect the certified uncertainty interval in such a way that their contribution is independent of the number of participating laboratories:

$$u_c(II) = \sqrt{\sum_{i=1}^{n} [u_i(II)]^2} \quad (7)$$

i = category II uncertainty identification number, varying from 1 to n
Typical examples of this category are the use of a common calibrant by all laboratories or material-related effects such as between-units variation (see "Effect of possible inhomogeneity and instability on the certified uncertainty").

3. Uncertainties in between the two above categories $[u_c(III)]$
These are *common to groups of limited numbers of laboratories* $\sum_{q=1}^{g} h_q = l$, such as those using an identical analysis procedure:

$$u_c(III) = \sqrt{\frac{\sum_{q=1}^{g} h_q \cdot [u_c(q)]^2}{g \cdot l}} \quad (8)$$

with:

$$u_c(q) = \sqrt{\sum_{i=1}^{n} [u_i(q)]^2} \quad (9)$$

q = group identification number, varying from 1 to g
g = total number of groups
l = total number of laboratories
h_q = number of laboratories in group q
i = category III uncertainty identification number in group q, varying from 1 to n

4. Moreover, as all laboratory means are not completely identical, a *residual component* $[u(R)]$ corresponding to the standard uncertainty of the average of the laboratory means should be considered as well:

$$u(R) = \frac{s_{betw}}{\sqrt{l}} \quad (10)$$

s_{betw} = standard deviation of the laboratory means
l = total number of laboratories

As already indicated, if the expanded uncertainty of each laboratory is correctly estimated, all laboratory results should overlap. More specifically, one can state that in fact laboratories within the same group should have mean values $x(j)$ differing from the mean group value $\bar{x}(q)$ by less than:

$$|\bar{x}(q) - x(j)| \le k \cdot u_c(I,j) \quad (11)$$

whereby $u_c(I,j)$ corresponds to the combined category I uncertainty of laboratory j, whereas all laboratory mean values $x(j)$ should differ from the overall mean value $\bar{x}$ by less than:

$$|\bar{x} - x(j)| \le k \cdot \sqrt{[u_c(I,j)]^2 + [u_c(III,q)]^2} \quad (12)$$

whereby $u_c(III,q)$ corresponds to the combined category III uncertainty of the group to which laboratory j belongs.

Laboratories whose results do not overlap within these limits are either affected by unrecognized systematic errors and/or by uncertainties that have been underestimated or omitted. Their results should therefore not be considered for certification.

The final uncertainty of the laboratory intercomparison can then be calculated as:

$$U = k \cdot \sqrt{[u_c(I)]^2 + [u_c(II)]^2 + [u_c(III)]^2 + [u(R)]^2} \quad (13)$$

Effect of possible inhomogeneity and instability on the certified uncertainty

Most frequently the between-units variability resulting from a homogeneity study is not insignificant compared to the uncertainty of the mean value. In addition it is generally preferred to assign a single certified value to all units of the entire CRM batch. Therefore, the uncertainty associated with the (possible) between-units inhomogeneity of the material should be included in the total uncertainty of the CRM. As indicated in [5], this can be done either by basing the CRM uncertainty on the statistical tolerance interval of the homogeneity study or by including the between-units standard uncertainty in the "category II" combined uncertainty $[u_c(II)]$ calculated according to Eq. 7.

The within-unit inhomogeneity, on the contrary, should in general not be included in the CRM uncertainty, except if such small sample intakes are used (e.g. in microanalysis techniques) that the sample inhomogeneity becomes significant compared to the certified uncertainty of the CRM. The main difference with between-units homogeneity testings is that if the observed within-unit inhomogeneity is significantly larger than the CRM uncertainty, it is sufficient to recommend the use of a larger sample intake on the basis of the fact that the uncertainty due to material inhomogeneity is inversely proportional to the square root of the mass of the analysed sample [6]. It is on the basis of this property that microanalysis was effectively proposed to determine experimentally the minimum sample mass down to which CRM certificates remain valid [7].

Linear regression and correlation can be used for the prediction of the possible instability of CRMs [8]. Quantitative characteristics expected to decrease (or

increase) with time are determined by calculating the time at which the 95% lower (or higher) confidence limit intersects the acceptable lower (or higher) specification limit, i.e. the lower or higher limit of the certified interval. The time so determined may then be considered as the expiration date, as one may be 95% confident that the average value of the batch characteristic will remain within specification until that date. As was the case for the within-unit variation, this possible instability should, in general, not be included in the CRM uncertainty, except if the degradation is significant compared to the certified uncertainty of the CRM. In such cases it might be preferred, rather than to reject the material as CRM, to certify an arbitrarily chosen interval within which the material can be expected to remain stable during a significant period of time, i.e. until the expiry date of the certificate.

Conclusion

The Guide to the expression of uncertainty in measurement provides a framework for assessing uncertainty which can and should be used for the certification of reference materials by laboratory intercomparison. However, as is stated in its paragraph 3.4.8., the following should be noted:

– *It cannot substitute for critical thinking, intellectual honesty, and professional skill.*

– *The evaluation of uncertainty is neither a routine task nor a purely mathematical one and depends on detailed knowledge of the nature of the measurand and of the measurement.*

– *The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value.*

This is particularly the case for the certification of reference materials. The above procedures can be used to obtain an estimation of both the certified value of a reference material and its uncertainty. However, there must be room for critical evaluation of the results by the people and organizations taking up responsibility for the values assigned to a CRM. Therefore it may be common practice in some organizations to increase the calculated uncertainty as it is felt to be optimistic. One should however be careful not to give lower uncertainties just on the basis of the fact that large uncertainty intervals may be interpreted as being the consequence of e.g. an analytical artefact.

References

1. Guidelines for the production and certification of BCR reference materials (1997) - document BCR/01/97, European Commission, Dg XII-5-C (SMT Programme).
2. Guide to the expression of uncertainty in measurement (1995) ISO, Geneva, ISBN 92-67-10188-9
3. Giacomo P (1987) Metrologia 24:49–50
4. Quantifying uncertainty in analytical measurement, 1st edn (1995) Eurachem, ISBN 0-948926-08-2
5. Pauwels J, Lamberty A, Schimmel H, Homogeneity testing of reference materials, Accred Qual Assur 2:51–55
6. Ingamells CO, Switzer P (1973) Talanta 20:547–568
7. Pauwels J, Vandecasteele C (1993) Fres J Anal Chem 345:121–123
8. Pauwels J, Lamberty A, Schimmel H, Quantification of the expected shelf-life of certified reference materials, Fres J Anal Chem (accepted)