Certification of matrix reference materials in chemical measurement: a new basis for adopting a consensus value derived from a proficiency test

Michael Thompson[a], Philip J Potts[b] and Peter C Webb[b]

ISO Guide 35 (2017) specifies conditions for the certification of matrix reference materials via information from proficiency tests. Many scientists involved in the provision of proficiency tests for analytical chemistry have concluded that the Guide 35 conditions are *effectively* fulfilled by identifying certified values with properly-derived consensus values. However, the claim that such reference materials are 'certified' is likely to be challenged by traditional metrologists. This paper proposes to move beyond this unresolved situation to put certification *via* proficiency testing on an unassailable footing. The proposal therefore is not a questionable fulfilment of the Guide 35 conditions but an unrelated paradigm for certification.

[a] School of Biological and Chemical Sciences, Birkbeck University of London, Malet Street, London WC1E 7HX, UK.

[b] Faculty of Science, Technology, Engineering and Mathematics, The Open University, Milton Keynes, MK7 6AA, UK.

Terminology as used in this paper

Consensus: a location estimate characterising the maximum density of an effectively unimodal dataset.

Standard error: the standard deviation of an estimated statistic (*e.g.*, a robust mean or mode of a dataset) as contrasted to that of a simple variable (*e.g.*, a result of a measurement).

Matrix: the constituents and partitions associated with the analyte in a test material, treated test material, or calibrator. (*Note*: a partition, for example, could be a discrete mineral phase of a rock, or seeds in a chopped whole plant.)

Invalid result: result of a measurement: (a) apparently expressed in incorrect units; and /or (b) obtained by an inappropriate measurement principle, method or procedure; and/or (c) of a value that can reasonably be regarded only as the outcome of a procedural blunder.

Scope of paper

The discussion is restricted to instances where the measurand is a mass fraction, mole fraction, or concentration estimated directly (i.e., not *via* an independently-varying proxy analyte), and the results are expressed on a ratio scale, with sufficient digit resolution to allow the statistical theory of continuous distributions to be applied without resultant problems. '*Results from a proficiency test*' and cognate phrases should be taken as referring to results for a single analyte derived from a single test material, in a single round of a proficiency testing scheme.

(*Note*: a proxy analyte is determined when the true analyte is not directly addressed. MS methods can exemplify this complication when the mass fraction of a single isotope is used to imply that of the naturally-occurring mixture. This could cause inaccuracies in instances where the natural isotopic ratios vary substantially.)

Introduction

Proficiency testing became widespread after the publication of the first edition of the ISO/IUPAC/AOAC Harmonised Protocol[1] and its recognition as an essential prerequisite for accreditation. Proficiency testing has since then given chemists access to an extraordinary amount of new information, well beyond that required solely for testing the performance of individual laboratories[2]. It was quickly perceived by proficiency test providers that this information was *inter alia* relevant to characterisation of matrix reference materials. As initially it had no role in the ISO REMCO[3,4,5] certification procedures, and sprang from a different philosophy of measurement, information from proficiency tests was at first regarded as impossible to incorporate into certification. The 2017 update of ISO Guide 35 (ref 3) has moved away from that position to a degree, but still imposes a number of conditions on the use of proficiency test information in certification, conditions that are, if taken at face value, impossible to fulfil given that the primary purpose and cost-efficiency of proficiency testing must not be subverted by the introduction of extraneous technical requirements[6].

Many scientists involved in proficiency test provision have concluded that the Guide 35 conditions for certification are *effectively* fulfilled by properly-derived consensus values from properly-conducted proficiency tests. However, the claim that such reference materials are 'certified' is likely to be challenged by traditional metrologists. The sticking point is the fact that 'certified reference material' is defined in VIM3 (Ref 7) in terms of 'property values having associated uncertainties and traceabilities obtained by using valid procedures'.

While a cogent case has been made that the VIM3 definitions can be effectively met *via* proficiency test results[8], the present paper seeks to move beyond this unresolved context and put certification *via* proficiency testing on an unassailable footing. This proposal, therefore, is not a questionable fulfilment of the Guide 35 (2017) conditions but an alternative view of certification.

Basic principles for an alternative route to certification

The following basic principles together provide a sound conceptual framework for characterising matrix reference materials for chemical measurement from the results of a sufficiently-large proficiency test. At first sight these principles seem counterfactual and at variance with commonly held perceptions in metrology. However, they cannot logically be dismissed out of hand simply because they differ from the metrological orthodoxy: they are substantiated in detail in the discussion below. This discussion follows the pattern: (a) what is a consensus? (b) how do you determine a consensus and its uncertainty? and (c) what are the metrological properties of a consensus?

The specific principles are as follows.

1. *Certified value*. There is quite generally no privileged way of estimating a certified value. The only *fundamental* requirement for certification is a demonstration of unbiased estimates of the value of the measurand itself and of its uncertainty.
2. *Common bias*. In an interlaboratory consensus, the common bias (as contrasted with the laboratory-specific or method-specific biases) is unknowable.
3. *Uncertainty*. An unbiased estimate of uncertainty in a consensus can be obtained as its robust standard error.

Other broad underlying principles are applied to the whole of the detailed discussion, as follows.

(a) All results of measurements are biased, and the same applies to the location estimates derived from them. The only point at issue is whether the knowable bias is small enough to be inconsequential[9].
(b) It is impossible to specify an inflexible statistical protocol that can deal adequately with any contingent dataset: the use of expert judgement is *essential* for determining the value of a consensus, for identifying a consensus as a certified value, or for abandoning the attempt.
(c) Traceability is not a valid basis for certifying consensus values derived from interlaboratory studies. The discussion shows, however, that a properly-considered consensus has an intrinsic validity and an uncertainty that does not stem from the notion of traceability.

What is a consensus?

In proficiency testing of chemical measurement, we take 'consensus' to mean that most of the results are in 'reasonable agreement', that is, no more disperse than usually encountered in chemical measurements under reproduced (interlaboratory) conditions. What is usually observed in such a dataset is a unique value of the measurand where the density of results is greatest. We can call that point the 'mode' if we think in terms of an underlying smooth population of results from which our discretised sample has been derived. (The simple definition of 'mode' as the most frequent value is unhelpful in this context.) If there clearly seems to be only one mode in a sufficiently large dataset, we are entitled to call that a 'consensus'. Of course, complications arise, because in most moderate-to-large datasets there will be a small proportion of results, referred to as 'outliers', that do not seem to belong to the main distribution. The presence of outliers in results from a proficiency test calls for appropriate handling.

In some datasets we see more than one range of the measurand where there is a significantly higher density of observations, in which case we refer to the datasets as 'multimodal'. When we encounter multimodality, we cannot nominate a particular mode as a consensus, unless there is irrefutable *additional* evidence (that is, other than the data themselves) that the supplementary modes are derived from invalid measurement procedures. There is a clear need for expert judgement here, because it is quite possible to regard a small random cluster of outliers (or even a single outlier) as a small mode. Furthermore, there is no satisfactory test for statistically significant multimodality.

Statistical approaches to finding a consensus

To quantify a consensus, we attempt to estimate the modal value of a smooth distribution that, in an optimal ('best-fitting') way, underlies the observed results. If a distribution appeared to be unimodal, close to symmetrical, and without outliers, the arithmetic mean would be the best estimator of the mode. However, such an outcome is rare in proficiency testing. More often, the distribution seems unimodal and close to symmetrical but includes a small proportion of outliers. In such instances a robust mean is nearly always the best estimator of a mode. (Robust statistical methods downweight results unduly far from the mode but where necessary compensate for the downweighting. The median is a simple type of robust mean.)

When the distribution of the dataset seems to be unimodal but skewed, a robust mean is usually unsuitable to be regarded as a consensus. The mode, however, is still intuitively the best consensus. Unfortunately, the mode cannot be reliably estimated by inserting the results into a formula. This difficulty arises because skewed continuous parametric distributions (such as the lognormal) are seldom found empirically to provide an adequate fit to proficiency test datasets. Methods for estimating modes non-parametrically seem not to have been studied in depth, but several methods have been proposed, all of them computer-intensive[10].

Uncertainty of a consensus

Standard uncertainty is identical with the standard deviation of an unbiased result or the standard error of an unbiased consensus. In a proficiency-test dataset showing a quasi-symmetrical distribution with outliers, a robust estimate of reproducibility standard deviation ($\hat{\sigma}$) is an appropriate starting point. It downweights observations far from the central

tendency and thus describes the variation of the 'sensible' results. There are several useful algorithms for calculating a robust standard deviation. The standard error of the consensus can then be calculated from it and safely identified as the uncertainty on the consensus. An unbiased estimate of the standard error would take the form $\hat{\sigma}/\sqrt{n^*}$, where n^* is the effective number of results, that is, the total number (n) adjusted for the downweighting in the robustification process. In instances where the proportion of outliers is small, the standard error can safely be taken as $\hat{\sigma}/\sqrt{n}$, which would give rise to a small underestimate. However, the underestimate is unlikely to exceed 5% of the total in practical instances[11].

If the consensus of a distribution is estimated non-parametrically, the estimation of its standard error directly by using a resampling procedure[12] (the bootstrap for example) is the only recourse available at present.

The standard error of a consensus takes account of *all* variation within the dataset, measurement variation occurring under conditions of reproducibility (between-laboratory) and of heterogeneity in the test material. No additional terms need to be added to the standard error of a consensus for it to comprise an unbiased estimate of its uncertainty.

Of course, a certified reference material has to be sufficiently close to stable, but that is an issue best considered separate from uncertainty. Tests for stability comprise a normal feature of a proficiency test. Uncertainties related to separate instability tests must not be combined with the uncertainty of the consensus evaluated as its robust standard error[13]: that follows because in instances where instability is null, the extra uncertainty arising from the test arises from the test alone rather than from the reference material.

Bias in a consensus

Simply forming a consensus from a dataset serves to differentiate between (a) a bias that is common to all the members of the dataset and that contributes to the consensus, and (b) residuals (unique deviations for each member of the dataset) that together form a random distribution characterised by a standard deviation. From our general principles, all consensus values will include a degree of common bias. To justify certification, the bias in the consensus should be negligible in relation to the uncertainty in the consensus. However, such bias is unknowable, because the analyst cannot know the true value to make the comparison. How then should we proceed?

In fact, there is every reason to expect the unknowable bias to be small in mature (stabilised) chemical proficiency tests. Well-known sources of bias in analytical procedures will have been eliminated or at least reduced to inconsequential levels during their development and validation. Moreover, most analytical procedures, even those based on distinct measurement principles, are structurally homologous. They have the same basic conceptual modules that are connected in the same way. Essentially: the treated test material and the calibrators are compared *via* a calibration function by their response on a common scale that is a surrogate for mass (light emitted or absorbed, number of ions collected, electrical potential, etcetera). Therefore, blunders aside, all such results should tend to give the same result, apart from individual biases, mostly derived from incomplete recovery, matrix-mismatch, and attempts to correct them. A consensus, therefore, goes well beyond a simple 'wisdom of crowds' idea. It would be very unlikely that numerous laboratories would arrive independently at estimates close to the same inappropriate value.

Before a consensus can be accepted for certification purposes, however, it must be shown to exhibit no signs of detectable bias. An essential preliminary therefore comprises a visual examination of various representations of the dataset. Obvious problems are manifested as incipient bimodality or strong skewness visible in a kernel density or in a display of the ordered results. An abnormally wide dispersion (outliers apart) in the dataset is also suspect: the robust standard deviation of the results must be consistent with the expected outcome for the relevant mass fraction of the analyte, the nature of the test material, and the measurement principles.

Suspected bias can sometimes be detected (or eliminated from consideration) in a set of proficiency test results by comparing the location estimates of independently-meaningful subsets of the data. Appropriate discriminating factors for such subsets could be *inter alia* the different measurement principles, analytical methods or decomposition procedures applied to the test portion. Bias would be indicated by significant differences between (or among) the separate consensuses of the subsets. Where such discrepancies are found, it may be impossible to determine a certified value.

When no such statistically-significant discrepancies are found between subsets, we must recognise that these latent differential biases are in principle still present, and possibly big enough to affect the value of a consensus. However, these biases will also automatically inflate the uncertainty estimated as a robust standard error of the whole set. This inflated uncertainty will tend to accommodate any concomitant bias and thus avoid misleading outcomes.

In the absence of the indications of bias outlined above, there is no way of addressing the common bias in a set of results from a proficiency test. In these circumstances, analysts have no alternative but to treat unknowable common bias as negligible. In some instances, of course, the participants, or a large subset of them, will be using a standard procedure, which is tantamount to a defining or empirical procedure[14]: in those circumstances a consensus will be bias-free by definition.

Bias with one measurement method

One special circumstance needs to be considered: proficiency test providers usually have no control over measurement procedures used by the participants. Possible outcomes of a proficiency test could include a set of results that are dominantly, or even exclusively, derived from a single measurement method. The problem implied is that consensus results could be affected to an important extent by undetectable bias in the single method. This worry can usually be readily dispelled. To do that, we have to consider the evolutionary history of methods used in a particular analytical application sector.

In an early state of affairs in an application sector, often there will be many competing analytical methods in use for the same task. Economic factors will ensure that, over time, less-fit methods will cease to be used[15]. The application sector will therefore change, without any overarching supervision, towards fit-for-purpose methods. This evolutionary process has the effect of reducing the variety of measurement principles being applied, in some cases to the point where there is only one method remaining in use in the sector. This happens because analysts (and, indirectly, their customers seeking best value for money) have decided that the remaining method is the best available. There will be in principle a residual bias in a one-method consensus but, in effect, the customers will be content with the state of affairs, because it is approaching fitness for purpose. Therefore, there is no special problem in a

mature application sector with defining a consensus from data overwhelmingly produced by one method. Indeed, if the method is regarded as 'standard' or self-defining, there will be no bias by definition.

'One-method' certification is appropriate, however, only in mature application sectors which have evolved from an earlier multiple-method situation. When an analyte becomes the focus of a proficiency test for the first time, there may well be only (or predominantly) one, relatively-untried, method for its determination. This might happen, for example, if the analyte were a food contaminant newly considered to be a health hazard. In such an instance there may well be undiscovered interference in the analytical method or procedure and corresponding bias in the results. In these circumstances, however, it would be premature to be thinking in terms of certifying a reference material.

Conclusions: Requirements for the acceptability of a certified value based on proficiency test results

A paradigm for the characterisation of matrix reference materials by using results from proficiency tests has been presented. A protocol for its acceptable execution would need to include constraints on the way in which it could be safely applied. To quote a skeptical colleague, 'You can't let unreliable organisations derive a certified value from any old dataset'. Prudence therefore suggests that, for each certified measurand, the following list of requirements is necessary and, details apart, sufficient. Note that it is considered detrimental to stipulate, in advance of the data, a single statistical protocol for determining assigned values and their uncertainties. On the contrary, a degree of expert judgement, applied within the bounds of the nine requirements below, seems *essential* to the process. Such judgement must, of course, be clearly justified.

1. *Qualifying proficiency testing scheme.* The proficiency test in question must be open to all-comers, mature, and widely recognised as compliant with the IUPAC Harmonised Protocol[16] or of equivalent status.

2. *Number of participants' results.* A sufficient number of valid results must be available. Fifteen valid results is often regarded as the necessary minimum number.

3. *Statistical approach.* The expert use of robust estimation or other methods of similar sophistication is essential.

4. *Consensus and uncertainty.* A clear consensus is essential.

5. *Bias.* There must be no observable bias among the results contributing to the consensus.

6. *Dispersion.* There must be no abnormal degree of dispersion among the valid results.

7. *Stability.* The reference material must be taken as either indefinitely stable or a use-by date defined for the consensus.

8. *Expert group.* There must be consideration of all relevant details by an expert group.

9. *Record keeping.* The expert group must keep and make available adequate records during the proposed lifetime of the reference material.

(*Notes.* The group of collaborating experts should all be scientists qualified to a professional status equivalent to the UK 'Chartered Chemist' designation. This group must comprise as separate individuals at least: (a) a person familiar with the types of materials being analysed, (b) an analytical chemist familiar with the strengths and shortcomings of the relevant analytical measurement principles, methods, and procedures, and (c) a person experienced in the statistical procedures used to determine the consensus value and its standard error. The records kept must include, for each certified value, details of: the nature and origin of the test material; the source of the data; the technical considerations and decisions made; and the experts' identities, qualifications and possible conflicts of interest.)

Epilogue

Some producers of reference materials may feel that the proposed conditions would be unduly open to inadvertent or deliberate malpractice. They should note that certification according to any procedure would be no less prone to these potential shortcomings—all approaches rely entirely on good professional conduct.

Although such comparisons are invidious, in some ways the present alternative paradigm of certification seems to approach fitness for purpose more surely than the ISO Guide 35: (2017) route: (a) it is less prone to bias as it usually engages a much greater number of self-elected laboratories using a range of currently-acceptable analytical methods under routine conditions; and (b) the uncertainty on the PT-based certified value is quite often the smaller, again by virtue of the large number of participant laboratories. Additionally, of course, it is cheaper to administer.

It is difficult to compare, in an incontestable way, certified values determined by the ISO Guide 35 route and the proficiency test route. Ideally a comparison should be conducted in a way such that the two outcomes are completely independent. That in turn requires the use of disjoint sets of laboratories for the ISO and PT routes, and double-blind organisation of the study. Attempts so far[17], show a very small proportion of statistically-significant differences between ISO-route certificate values and proficiency test consensus values. Although reasonably convincing, these studies have not managed to achieve complete independence.

Conflicts of interest

The authors have no conflicts of interest to report.

[1] M Thompson and R Wood. *Pure Appl. Chem.* 1993, **65**, 2123-2144.
[2] Analytical Methods Committee. *Accredit. Qual. Assur*. 2010, 15, 73-79.
[3] ISO Guide 35 (1989)
[4] ISO Guide 35 (2006)
[5] ISO Guide 35 (2017)

[6] Analytical Methods Committee. *Accredit. Qual. Assur*., 2010, **15**, 73-79.
[7] Bureau International des Poids et Mesures. *International vocabulary of metrology—basic and general concepts and associated terms (VIM)* (3rd Edition). JGCM 200:2012.
[8] P J Potts, P C Webb and M Thompson. *Geostandards and Geoanalytical Research*, 2019 (Accepted/In press)
[9] Ramsey M. H. Thompson M. and Banerjee E. K. *Analytical Proc*. 1987, **24**, 260-265.
[10] M Thompson. *Analytical Methods*, 2017, **9**, 5534-5540.
[11] M Thompson. *Accred. Qual. Assur*. 2006, **10**, 574.
[12] B Efron and R J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York, 1993.
[13] Analytical Methods Committee. *AMC Technical Briefs* No. 61 in *Analytical Methods*, 2014, **6**, 3201-3202.
[14] M Thompson. *Analytical Methods*, 2016, 8, 4908-4911.
[15] M Thompson. *Analyst* 1999, **124**. 991.
[16] M. Thompson, S. L. R. Ellison, and R. Wood. *Pure Appl. Chem*. 2006, **78**, 145-196.
[17] P J Potts, M Thompson and P C Webb. *Geostandards and Geoanalytical Research*, 2015, **39**, 407-417.