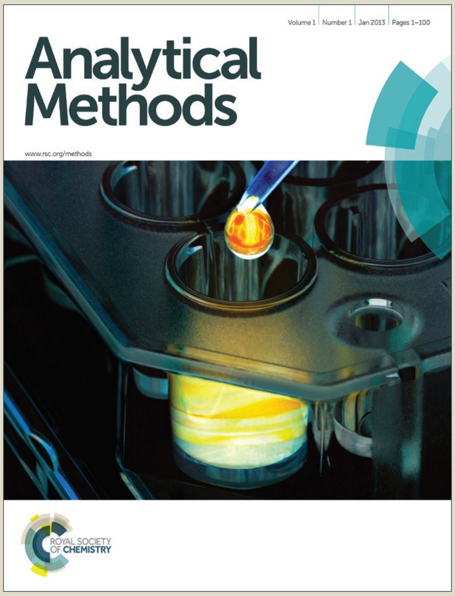
Analytical Methods

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analytical Methods
www.rsc.org/methods
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

On matrix reference materials characterised by proficiency test

Michael Thompson
School of Biological and Chemical Sciences
Birkbeck University of London
Malet Street
London WC1E 7HX, UK.
m.thompson@bbk.ac.uk

Abstract

This paper examines the status of a matrix reference material characterised by the consensus of the results from a proficiency test. It is shown that the very existence of a consensus in chemical measurement attests to its validity. Unique biases in individual laboratories are largely averaged out in forming the consensus. All of the variation between laboratories' results (including that resulting from the unique biases) will be encompassed by the standard error of the consensus. Any common bias among the laboratories remains unknowable. The supposed absence of traceability in a consensus can be overcome by obtaining appropriate information from the selected participants. More fundamentally it is show that, in any event, a traceability chain is indeed broken in many types of chemical measurement, without detriment to the validity of the result. The overall conclusion is that, *with carefully considered safeguards*, characterisation of matrix reference materials by proficiency testing is appropriate to establish the concentrations of selected analytes.

Introduction

Proficiency tests are designed principally to enable participant laboratories to detect and remedy shortcomings in their procedures[1,2]. Nevertheless, other quality-related activities such as method validation[3] exploit the results of proficiency tests. Currently in chemical measurement there is interest in the use of proficiency test results for the characterisation of matrix reference materials, with the characterised value identified as the consensus of the participants' results and its standard uncertainty as the standard error of the consensus, both determined by appropriate statistical methods. Several proficiency test providers are currently issuing their surplus materials on the basis of this type of characterisation, so the time is ripe for a detailed examination of the practice. It is the theme of this paper that such a characterisation, if carried out with due consideration, can provide a stand-alone attestation of the material's composition, thus providing a valid and valuable addition to the analyst's armoury alongside certified reference materials.

In order to make this case, it is necessary to show that such characterisation is metrologically sound. Much has been written about how that might be achieved[4]. But fundamentally, metrological soundness does not depend on following a particular documented procedure: it implies only that there exists for a reference material a characterised value, with an associated uncertainty that is realistically estimated for a suitable mass of test portion. An analyst, armed with that information alone, could decide whether or not a reference material of appropriate matrix would be suitable for a particular purpose. Consequently, any approach whatsoever that provides that information is sufficient for characterisation.

It will be shown here that many proficiency tests can, with little or no adjustment, be made to conform to this fundamental requirement. The issue of traceability is often raised in relation to the status of a consensus, but a good case can be made that the apparent problems are insubstantial.

Location, bias and uncertainty

For present purposes we are interested in the 'location' of results from many laboratories participating in a round of a proficiency test. (A 'location', for example a mean, robust mean, median or mode, quantifies the tendency of a set of values.) A participant consensus used as the assigned value in a proficiency test is a location estimate and as such has a standard error. (A 'standard error' is the standard deviation of a statistic (such as a robust mean) as opposed to that of a simple variable (such as a result).) If a location estimate is unbiased, its standard error is by definition equal to its standard uncertainty.

How could we tell whether a consensus is unbiased? Well fundamentally we cannot! In principle all results of measurements of a quantity are biased, and so are locations of sets of results. Moreover, we cannot directly quantify the bias stemming from a specific procedure applied to a matrix material because the estimate is the difference between the location of a set of measurement results and the unknowable true value of the quantity. The key issue is whether the putative bias seems acceptably small. We can, however, estimate the difference between the locations of results obtained by applying two distinct analytical procedures to portions of the same material. From the comparison we can then refine the procedures in ways that we think will reduce bias, according to our knowledge of the physical and chemical principles involved. We tend to do that, using different methods and measurement principles, until we consider any residual bias to be acceptable.

In considering bias in a consensus, it is useful to break down an individual analytical result into distinct terms as follows: (a) the (unknown) true value; (b) a residual bias common to all participant laboratories; (c) the random variation in the individual laboratory; and (d) a unique bias within the individual laboratory. This is shown schematically in Fig 1 (for an unrealistically small number of laboratories).Both the within-laboratory random variation and the individual laboratory biases can, from the viewpoint of the whole dataset (our present concern), be treated as zero-mean random processes. When we have a long list of results for a particular determination (as in a proficiency test) these variations will be to a considerable degree ‘averaged-out’ in the consensus and be represented by its much smaller standard error. This standard error encompasses all sources of variation in the dataset, be they overt or latent. Only the common bias is present unchanged in the consensus and, as we have seen, the magnitude of that in principle remains unknowable.

Handling unknowable bias

It is proposed here that, *given reasonable safeguards*, this unknowable common bias must be treated as zero for all practical purposes. The safeguards amount to eliminating all seemingly relevant sources of bias during method validation. Having done that, if subsets of results based on a variety of physical principles converge on a single location (obviously within limits defined by the relevant uncertainties) we have no grounds to suspect bias unless we detect evidence or harbour well-founded suspicions to the contrary. Evidence would be manifested in the form of systematic discrepancy between locations of subsets the data associated with particular analytical procedures or methods. Suspicions would be based on professional experience of the analytical methodology and of the test material, reinforced by the appearance of a strong skew or multimodality in a dataset. Where no evidence or suspicions are forthcoming, we can say *nothing* about the magnitude of any residual bias. However, it is noteworthy that instances of latent bias, either common or unique, are most often discovered *via* the very process of proficiency testing.

Any characterisation procedure (including certification *via* ISO Guide 35[4]) that is based on interlaboratory study is subject to exactly these same problems as proficiency testing. In all cases, known bias is reduced by refining the analytical procedures as far as economically possible, then simply accepting that remaining. The only differences arise in factors affecting the selection of the subset of results regarded as ‘valid’ for characterisation purposes. It is usual in certification studies to attempt to test for bias in the results of individual participants. (This is discussed further under ‘Traceability’.) That intervention does not feature in proficiency testing. There is moreover in the certification of a reference material the capacity for referring discrepant results back to the originators with the possibility of revising them. (In principle that practice introduces a new ‘expectation bias’: if the revised suspect results are further away from the putative consensus, they will be ignored; if they are notably closer they will be included in forming a new consensus.) In proficiency test results, there is no possibility of retrospectively correcting biases in particular results or arising from particular analytical procedures.

A further difficulty is that the number of different measurement principles and methods in use tends to decline with time as certain methods and procedures are taken to be superior and become preferred by an increasing proportion of participants. This process mostly tends to reduce bias. But ultimately, when only one procedure remains in use, any residual bias will be untestable. This latent difficulty also affects all methods of characterising matrix reference materials based on interlaboratory study.

Traceability

A lack of traceability is often claimed (usually without detailed justification) to be a fundamental shortcoming of consensus values derived from proficiency test datasets. This assertion perhaps stems from an unwarranted distinction between proficiency test assigned values and values certified by interlaboratory study. Scrutiny, however, shows that the two procedures are homologous in their essential respects. Both procedures are based on a participant consensus. Both procedures make an estimate of the location of the results and its uncertainty from the results alone. The most striking difference between them is that certification usually requires each participant to justify the claim of a traceable result, while proficiency testing in the normal course of events does not. This distinction is largely insubstantial but, in any event, could be simply ameliorated.

This assertion rests on several points.

1. An error of consequential magnitude in a chemical measurement usually springs from two sources, (a) loss (or gain) of analyte during the process of chemical treatment of the test portion, and/or (b) loss or gain of net analytical signal brought about by matrix-mismatch between calibrators and treated test solution. It is conceptually incorrect and futile in practice to shoehorn these two effects into a traceability model because, contrary to prevailing doctrine, traceability is incomplete at these two points in a great majority of analytical procedures[5].

   Analysts are of course aware of these effects and moderate them in various ways, by correcting for recovery or reducing matrix-mismatch. But they seldom know how effective these attempts are in everyday analysis. Recovery, even when corrected, always differs from 100% by an unknown amount: matrices are always mismatched by an unknown amount. Sometimes these discrepancies are non-trivial. The traceability model therefore cannot help in estimating the often dominant uncertainty contributions brought about by these inescapable features of chemical analysis. This is *inter alia* why traceability-based models of an analytical procedure consistently tend to underestimate combined uncertainty, even in international key comparisons[6], and why, in properly designed studies, repeatability dispersion is virtually always smaller than (interlaboratory) reproducibility dispersion[7].

2. We can examine how a demonstration of traceability in certification is conducted and see whether that is rigorous. Certification bodies test for traceability typically by requiring participants in a study to analyse an existing certified reference material (CRM) alongside the candidate reference material and obtain a ‘satisfactory’ result. However, such a test is neither demanding nor conclusive. It would tend to have a low statistical power—typically a really large discrepancy would be needed to reach statistical significance. In addition, we could seldom guarantee that the candidate reference material and the CRM were effectively matrix-matched. Furthermore, how could we be sure that the certificated value of the reference CRM was itself unbiased? By harking back to successively older (and presumably less-rigorously certified) CRMs in a long regress back to uncertified calibrators? It would be easy for a proficiency test provider to ask participants to provide a comparable degree of reassurance on this point. For instance, participants could be asked to confirm that they have (a) used traceably-calibrated equipment, (b) matrix-matched calibrators prepared from elements or stoichiometric compounds of known purity or from other matrix reference materials, (c) a properly validated procedure, and (d) internal quality control, and furthermore (e) participate in an appropriate proficiency test and

(f) operate in a well-found laboratory operating under appropriate infrastructure features. These conditions would be fulfilled in accredited laboratories, so a statement of accreditation would suffice. (Obviously the typical occurrence of outlying results suggests that a minority of proficiency test participants do not conform to this specification, but such results can be managed by use of common sense and appropriate statistical procedures.) It is an interesting sidelight on this question to note that no 'demonstration' of traceability is suggested as necessary in the use of proficiency test data in method validation[3].

3. However, even this documentary reassurance may not be strictly necessary, as the very existence of a clear consensus in proficiency test results attests to its appropriate value. Only two circumstances can account for a clear consensus: either (a) the consensus is effectively unbiased, or (b) there exists an unknown common bias (that is, a bias affecting equally each of the participants' results). Laboratory-specific biases are largely averaged out in the consensus. The location estimate effectively includes only the unknown common bias but that, according to the previous argument, will usually have been reduced to a negligible level during the development/validation of the analytical procedures.

Uncertainty of the consensus

The standard error of the proficiency test location encompasses the effects of *all* causes of variation among participants, be they overt or latent. To equate a standard error of a consensus with a standard uncertainty implies that there must be no *known* common bias among participants' results, otherwise the fundamental idea of uncertainty—that all known sources of bias in the measurement procedure have been removed—is invalidated. Accepting that no known bias is present, it is clear that the standard error of the location estimate is indeed its standard uncertainty. It encompasses *all* variation at the time of the test, including heterogeneity. This identification is sometimes regarded with scepticism by scientists who follow established certification procedures because the value of the standard error often seems to be unexpectedly small. However, that happens partly because the effective number of results available in a proficiency test dataset is usually greater than that economically feasible in certification procedures. The standard error of the location estimate is therefore tantamount to a standard uncertainty at the time of the test. Smallness of uncertainty is of course an invaluable asset in a reference material.

A note on heterogeneity and instability

CRM producers include in the uncertainty budget terms relating to heterogeneity and long-term instability. This approach is correct in principle, but there are practical shortcomings in adequately implementing the idea. The high cost of certifying a reference material ensures that only materials expected to have a long shelf-life are likely to be candidates. This in turn means that affordable tests for deterioration by direct determination of the analyte will tend to be ineffectual when the reference material is stored under normal working conditions: such tests will usually have low statistical power (be unable to demonstrate relevant levels of instability) and will provide estimated uncertainty contributions that are unreliable[8]. Producers are warned against the possibility of underestimation[9] under these circumstances but, remarkably, not against over-estimation. The latter is the more likely eventual outcome because estimates of zero will be common but will be discounted in favour of higher values. That feature could contribute to inflated uncertainty estimates on certified values.

Proficiency testing *per se* does not rely on long-term stability of the test materials, whereas that is a key property of certified reference materials. However, an alternative strategy to preliminary instability tests is long-term monitoring of the material. Of course, the issuing body, at the time when the reference material was released, would have to be convinced that the material was going to be effectively stable, by virtue of prior experience with similar materials. Deterioration if any could then be followed over an extended period. Proficiency test providers, however, are in a uniquely powerful position to do that, simply by a participant-blind re-issue of the same material at future dates and a comparison of the consensus values[10]. The use of consensus values derived from a large body of participants obviates the problem of analytical run bias that potentially confounds results from a single laboratory repeated after an interval of time. (A valuable corollary of this idea is that, for stable reference materials, the consensus itself is shown to be stable.)

Conclusions

A substantial case has been presented for the recognition of matrix reference materials, characterised by the judicious consideration of proficiency test results, as stand-alone exemplars of chemical composition. The characterised value for an analyte and its standard uncertainty would be identified as the assigned value and its standard error. Such materials would comprise an invaluable and relatively inexpensive supplement to certified reference materials in the analyst's toolkit. It seems advisable for the moment to maintain a distinction between 'characterised reference materials' and certified reference materials, at least until a large body of experience in their preparation and use has accumulated.

The much-raised question of traceability in a consensus could be accommodated where necessary simply by a relevant statement from each qualifying participant. However, in most instances the very existence of a clear consensus is evidence of an appropriate value. This follows because in validated procedures any remaining bias must perforce be taken to be of negligible magnitude. Finally it has been shown that, in any event, metrological traceability is usually unavoidably incomplete in chemical measurement without detriment to the validity of the result.

Of course, a bald proficiency test consensus with no further support would be grossly insufficient to comprise 'characterisation'. A detailed documented study of each dataset would be required, involving expert consideration of the test material, the analytical procedures used, the statistical methodology and, indeed, the results themselves. Oversight of characterisation should be provided by accreditation agencies. A short informal guide to good practice would be a helpful development for this purpose.

[1] The International Harmonised Protocol for the proficiency testing of analytical chemistry laboratories. *Pure Appl. Chem*. 2006, **78**, 145-196.

[2] ISO 17043:2010. *Conformity assessment—general requirements for proficiency testing*.

[3] Analytical Methods Committee. *Accred. Qual. Assur*. 2010, **15**, 73-79.

[4] ISO Guide 35:2006. *Reference materials—general and statistical principles for certification.*

[5] M. Thompson. *Anal. Methods*, 2016, **8**, 940-941.

[6] M. Thompson and S.L.R. Ellison. *Accred. Qual. Assur*. 2011, **16**, 483-487.

[7] M. Thompson and R. Wood. *Anal. Methods*, 2015, **7**, 375-379.

[8] M. Thompson. *Anal. Methods*, 2015, **7**, 1627-1629.

[9] ISO Guide 35 paragraph 7.3

[10] M. Sykes and M. Thompson. *Anal. Methods*, 2015, **7**, 9753-9755.

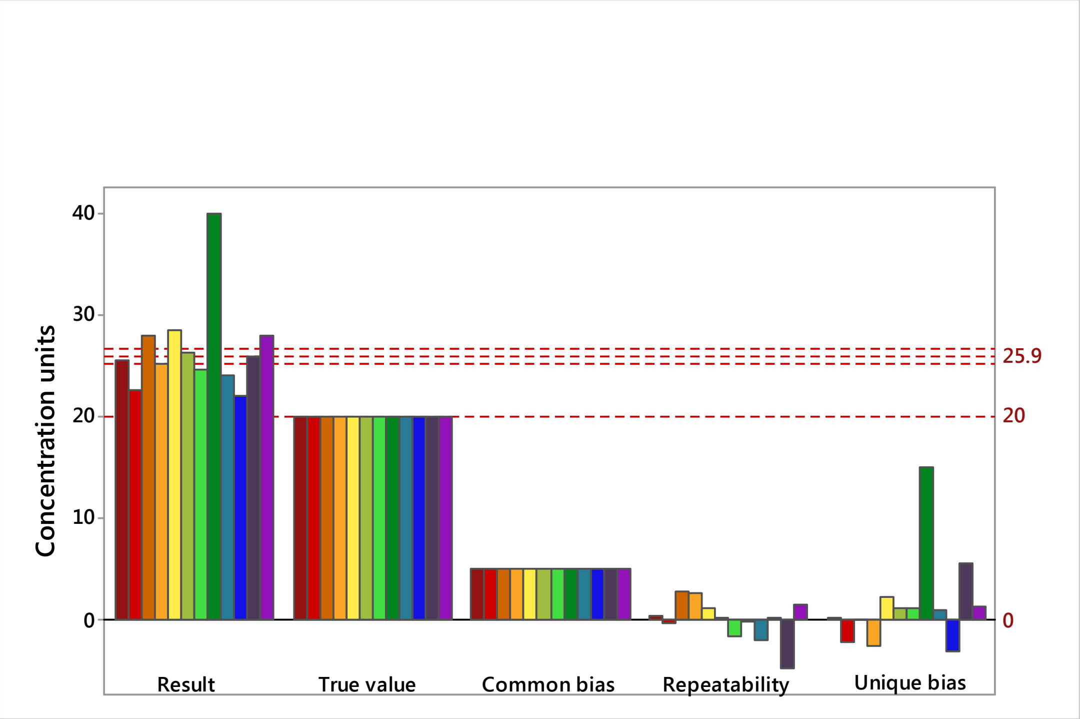
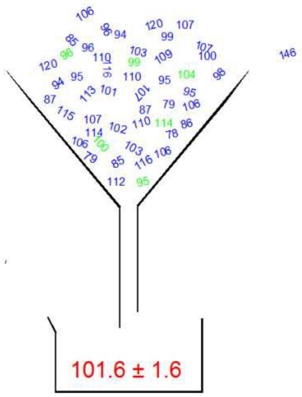


Fig 1. Schematic breakdown of analytical results (bars) in a proficiency test, into the true value (20), the common bias (5), the zero-mean repeatability variation and zero-mean unique bias. Individual participant laboratories coded by colour. The upper reference lines show the robust mean (a consensus) and its 95% confidence limits. The common bias (shown relatively large here for clarity) is included in the consensus.

A properly-determined consensus from a proficiency test is a metrologically-sound indication of chemical composition.


106
96 94
120 107
99
85 96
120 96 110
103
99 109
107 100
146
116
94 95 110 95 104 98
87 113 101
101
95
87 79 106
115 107 102 110 114 86
114 78
106 100 103
79 85 116 106
112 95
101.6 ± 1.6