Do we really need to account for run bias when producing analytical results with stated uncertainty? Comment on 'Treatment of bias in estimating measurement uncertainty' by G. E. O'Donnell and D. B. Hibbert

Rouvim Kadis*

Treatment of bias is an important issue relating to analytical quality. Recently, G. E. O'Donnell and D. B. Hibbert (*Analyst*, 2005, **130**, 721) recommended to always correct analytical results for 'run bias' determined by a single analysis of a certified reference material (CRM) in each analytical run. In the authors' opinion, this is necessary for the results obtained to be comparable from run to run. It is argued here that such a recommendation is logically inconsistent and stems from misinterpretation of measurement uncertainty as being estimated under repeatability conditions. The fundamental principle underlying the measurement uncertainty methodology is that all relevant sources of error should be taken into account, which results in overall uncertainty assessment and thus provides a means for a global comparability of measurement and test results. The local, *i.e.* run-to-run, comparability is not a factor if analytical results are interpreted on the basis of their associated uncertainty.

The assessment and treatment of experimental bias is an important issue relating to analytical quality. In a paper published in this journal, O'Donnell and Hibbert[1] address this issue with respect to measurement uncertainty estimation, since a generally accepted solution to the problem of bias handling in estimating uncertainty has not been found so far.

Two actions are generally feasible when analytical bias is detected: applying a correction for a (significant) bias, with the uncertainty of this correction added in the budget; or, alternatively, increasing the resulting uncertainty to take uncorrected bias into account. Currently, different approaches to account for uncorrected bias have been proposed[2–4] and the authors,[1] by applying Monte Carlo simulation to assess their performance, come to the conclusion that the best strategy is always to correct the results for bias, but when no correction is applied, the best way of increasing the uncertainty is the so-called SUMU_{Max} method,[2] that is summing up the absolute value of the bias and the value of expanded uncertainty.

Remarkably, the paper calls attention specifically to 'run bias'. An argument the authors bring forward is as follows: 'If we take the viewpoint of a client of a commercial laboratory who submits samples regularly to that laboratory for analysis, it is logical that the client would require these results to be comparable from run to run'. This leads them to the assertion that metrological traceability and therefore comparability of test results can be achieved if a correction for the run bias is carried out.

Accordingly, it is recommended to determine the run bias, δ_{run}, by a single analysis of a certified reference material (CRM) inserted in each batch. The observed deviation from the certified value of the CRM is first tested for statistical significance by comparing with its expanded uncertainty $U = ku(\delta_{run})$ where the coverage factor k is usually taken as 2 and the standard uncertainty $u(\delta_{run})$ is calculated as [eqn (7) in ref. 1]

$$u(\delta_{run}) = \sqrt{s_{p,r,CRM}^2 + u^2(C_{CRM})} \qquad (1)$$

The term $s_{p,r,CRM}$ represents here the pooled repeatability standard deviation of the CRM analyses, taken from the method validation studies, and $u(C_{CRM})$ is the standard uncertainty of the value carried by the CRM.

According to the authors' recommendation to correct the test result for run bias (whether it is significant or not), the value of δ_{run} is subtracted from each observed value, and the combined uncertainty u_c of the result is calculated from the formula [eqn (10) in ref. 1]

$$u_c = \sqrt{\frac{s_{p,r,Sample}^2}{n} + u^2(\delta_{run})} \qquad (2)$$

where $s_{p,r,Sample}$ should be read as the pooled repeatability standard deviation for a given type of sample, with n equal to the number of repeat determinations within a run made for the result reported.

Some remarks need to be made with regard to this recommendation.

First: generally, it is not a common procedure to analyze a CRM in each analytical run as the procedure described requires. Such practice may be associated with a limited stability of the analytical system when it is justified by the necessity to compensate for lack of control. In that case, however, the CRM measurement made within a particular run can be used for the direct comparison of the responses in a compensatory technique similar to the use of 'leader standard samples' mentioned by Kaiser[5] in his classification of calibration methods. This is a well-known technique that allows eliminating certain sources of variation in analytical measurement.

D. I. Mendeleyev Institute for Metrology, 19 Moskovsky pr., 190005 St. Petersburg, Russia. E-mail: rkadis@mail.rcom.ru; Fax: +7-812-323-9603; Tel: +7-812-323-9372

On the other hand, applying a correction is a calibration too, which brings up the question: why do we first need to obtain 'raw' results on the basis of a fixed calibration and then correct them for a bias? One cannot but see a logical inconsistency in using the two calibration methods in tandem: the second one might render the first essentially unnecessary.

It is pertinent to note that the analysis of reference materials to monitor accuracy (besides duplication to monitor precision) is a rational strategy of internal quality control (IQC) in analytical laboratories. Accordingly, the results of control analyses are indicative of the performance of the analytical system, but correction of analytical data on the basis of control analyses is regarded as unacceptable in the IQC guidelines.[6]

Second: there is a misinterpretation of what is behind the term 'run bias', δ_{run}, and how it is tested for statistical significance in the procedure suggested. The observed difference between the average of a small number of individual measurements or a single measurement on the CRM and its certified value [see eqn (4) and eqn (6) in ref. 1], apart from a random error contribution, is always a combination of run effect, laboratory effect, and method bias. Thus the observed difference is not either run bias or 'the actual bias of a particular analytical run' as the authors put it, but a total measurement error estimated on the CRM measurement.

In testing whether the observed difference is statistically significant the repeatability standard deviation is used by the authors on the ostensible ground that 'bias is often hidden by the significance test being based on reproducibility data'. Meanwhile, the available guidelines[7,8] on the use of reference materials (see also ref. 9) clearly demonstrate that checking of trueness is to be based on reproducibility, or intermediate precision data if the check is performed by only one laboratory. The tendency to find what is 'hidden' will evidently lead to falsely concluding that a significant bias occurs at a given level of confidence; hence, the need for a corrective action with regard to 'bias' is erroneously exaggerated.

What can we expect from the correction for 'run bias'? All the systematic error components cancel out, yet an additional source of uncertainty connected to the certified value of the CRM appears; besides, the random error contribution will be increased unless doubled [to make it sure, substitute eqn (1) into eqn (2)]. Furthermore, not taken into account in eqn (2), another uncertainty component will result from a mismatch (in composition and relevant properties) between the test material and the reference material. The CRM should closely resemble the unknown samples in relation to both the analyte content and the matrix – a prerequisite that cannot be fulfilled rigorously. Thus, the effect of cancelling bias is largely reduced or eliminated by an additional error introduced in the final measurement result.

Third: the fundamental principle underlying the measurement uncertainty methodology,[10] specifically as applied to chemical analysis,[11–13] is that all sources of error that are relevant should be taken into account. In other words, an overall uncertainty assessment is implied. If we refer to a hierarchical 'ladder of errors'[14] that are always present in chemical analysis, namely, repeatability error, run bias, laboratory bias, and method bias, it is easy to arrive at the following expression for the combined standard uncertainty u_c of a measurement[12]

$$u_c = \sqrt{\sigma_\varepsilon^2 + \sigma_{run}^2 + \sigma_{lab}^2 + u_b^2} \qquad (3)$$

where σ_ε^2 is the variance in repeatability conditions, σ_{run}^2 and σ_{lab}^2 are the variances of the populations of run and laboratory bias respectively, and u_b is the standard uncertainty for the method bias that can be estimated from a collaborative study. Eqn (3) represents the uncertainty in terms of a 'top-down' approach, with the additional contributions not covered by the error model[14] still needing to be taken into account.

Such overall uncertainty assessment conveys the true (metrological) sense of this concept. It is violated when one tries to use this concept where only local consistency of results is required, for instance, if a small difference between two test items needs to be detected. The term 'conditional uncertainty', proposed[12] for these cases as well as the so-called operational definitions of uncertainty[15] with different concept content that depends on specific conditions, is an attempt to adapt the measurement uncertainty methodology to the problem of a local comparative testing. However, this methodology is intended for a different purpose, and reducing the uncertainty to that in repeatability conditions, as in eqn (2), is an example of how the concept of measurement uncertainty should not be used. It is that very case, once more criticized recently,[16] "where the stated measurement uncertainty erroneously means 'repeatability' or 'reproducibility' (represented by a small so-called 'error bar'), whereas repeatability and reproducibility are only *components* of measurement uncertainty…".

The point is that with metrological concepts such as measurement uncertainty and traceability, the comparability of measurement results on the global scene (*i.e.* the inter-laboratory, international, or even 'intercontinental' comparability) is aimed to achieve. It is this global comparability that is pursued, because measurements of the same quantity made in different laboratories, at different times, and irrespective of the analytical methods used are to be comparable within their uncertainties in view of the need for their mutual recognition.[17,18]

If an internal consistency of results is only aimed at, an appropriate precision measure such as the repeatability limit is used and applying to the measurement uncertainty concept is not needed. In contrast, the run-to-run comparability the authors are striving for is not a factor if measurement and test results are interpreted on the basis of the stated uncertainty associated with the results, and any corrective action with regard to run bias or the thing called so by the authors proves unnecessary.

References

1 G. E. O'Donnell and D. B. Hibbert, *Analyst*, 2005, **130**, 721.
2 S. D. Phillips, K. R. Eberhardt and B. Parry, *J. Res. Natl. Inst. Stand. Technol.*, 1997, **102**, 577.
3 V. J. Barwick and S. L. R. Ellison, *Analyst*, 1999, **124**, 981.
4 V. Synek, *Talanta*, 2005, **65**, 829.
5 H. Kaiser, *Spectrochim. Acta, Part B*, 1978, **33**, 551.
6 M. Thompson and R. Wood, *Pure Appl. Chem.*, 1995, **67**, 649.

7 ISO Guide 33:1989, *Uses of certified reference materials*, International Organization for Standardization, Geneva, Switzerland.
8 R. Walker and I. Lumley, *TrAC, Trends Anal. Chem.*, 1999, **18**, 594.
9 T. Linsinger, *Comparison of a measurement result with the certified value*, European Reference Materials, July 2005 (http://www.irmm.jrc.be/html/reference_materials_catalogue/user_support/erm_application_notes/application_note_1/application_note_1_english_en.pdf; accessed 28 September 2007).
10 BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, *Guide to the expression of uncertainty in measurement*, International Organization for Standardization, Geneva, Switzerland, 1993.
11 EURACHEM/CITAC Guide: *Quantifying uncertainty in analytical measurement*, EURACHEM, 2nd edn, 2000.
12 Analytical Methods Committee, *Analyst*, 1995, **120**, 2303.
13 V. J. Barwick and S. L. R. Ellison, *Protocol for uncertainty evaluation from validation data*, report number LGC/VAM/1998/088, LGC Ltd., 2000.
14 M. Thompson, *Analyst*, 2000, **125**, 2020.
15 E. Hund, D. L. Massart and J. Smeyers-Verbeke, *TrAC, Trends Anal. Chem.*, 2001, **20**, 394.
16 P. De Bièvre, *Accred. Qual. Assur.*, 2006, **11**, 597.
17 W. Richter, *Fresenius' J. Anal. Chem.*, 1999, **365**, 569.
18 T. Quinn, *Measurement and Society with particular reference to metrology in chemistry, biology and medicine*, 2004 (http://www.inti.gov.ar/novedades/Presentaciones/01.pdf; accessed 28 September 2007).