Analytical Methods

PAPER

Cite this: *Anal. Methods*, 2016, **8**, 742

The comparison between reproducibility standard deviations from collaborative trials and proficiency tests: a preliminary study from food analysis

Michael Thompson

Received 2nd November 2015
Accepted 2nd December 2015

DOI: 10.1039/c5ay02890f

www.rsc.org/methods

Reproducibility conditions of the replication of a measurement include several circumstances. In chemical measurement 'reproducibility' is mostly taken to refer to an interlaboratory study, either a collaborative trial (that is, with a strictly defined analytical procedure) or a proficiency test (with no prescribed procedure). At first sight, we might expect the reproducibility standard deviation of the proficiency test to be the greater for the same determination: the various procedures used by the participants will each introduce an extra uncertainty related to their specific biases. No comprehensive study of this potential disparity has been undertaken hitherto. The issue is important because reproducibility standard deviation is closely related to standard uncertainty. Here a comparison is made between the trend of collaborative trial outcomes (standard deviation as a function of concentration) and individual values from the FAPAS proficiency testing scheme in the food analysis sector. Contrary to expectations, the general tendency is for proficiency tests to provide slightly smaller standard deviations than do collaborative trials at mass fractions of the analyte greater than 10^{-7}, and slightly higher at lower concentrations. However, there is considerable variation around the median level of the ratio at all mass fractions.

Reproducibility conditions for the replication of a measurement as currently defined[1] include several distinct circumstances but, in chemical measurement, are usually taken to refer to results from inter-laboratory studies, specifically collaborative trials and proficiency tests.[2] In a collaborative trial (or method performance study), selected laboratories analyse the same suite of test materials while using a single carefully-defined analytical procedure.[3] The reproducibility standard deviation is the between-laboratory value derived from the results by robust statistics or an equivalent outlier-rejection procedure. The test materials are selected from a single class of matrix and usually contain a range of concentrations of the analyte. In a proficiency test, however, laboratories are usually free to use any analytical procedure or method that seems appropriate. *A priori* it would be reasonable to expect robust reproducibility standard deviations derived from proficiency tests to be somewhat greater for the same measurand, because of the extra sources of variation introduced by the individual biases of different methods and procedures.

To test that expectation, the obvious approach *prima facie* would be to compare the outcomes of one or more collaborative trials and a number of rounds of a proficiency test, all dealing with a single analyte/matrix combination. A large number of such comparisons should enable the scientist to draw general conclusions. However, each individual comparison between the two types of interlaboratory study would be a laborious enterprise and not guaranteed to produce a clear outcome. Both tests would provide precision statistics at discrete but different concentrations of the analyte. That implies that the comparison would have to be carried out between models of precision (that is, standard deviation as a function of concentration) rather than between the individual values. This in turn raises the difficulty that collaborative trials are conducted with statistically-small numbers of laboratories, seldom greater than 12: the resultant standard deviations therefore would have wide confidence intervals and might give rise, without significant lack of fit, to a variety of possible models, some inappropriate.[4] Proficiency tests are less prone to this problem because the number of participant laboratories is usually considerably greater and the resultant standard deviations correspondingly more precise and, given enough time, more numerous. There is the additional complication that the ratio between the trends of the standard deviations might vary strongly with concentration. Finally, in either collaborative trials or proficiency tests, the precision statistics would sometimes be difficult to accommodate in an appropriate model: if the defined class of test material is too inclusive, "soil" for example, matrix effects can give rise to lack of fit. An example showing some of these difficulties is shown in Fig. 1.

School of Biological and Chemical Sciences, Birkbeck University of London, Malet Street, London WC1E 7HX, UK. E-mail: m.thompson@bbk.ac.uk

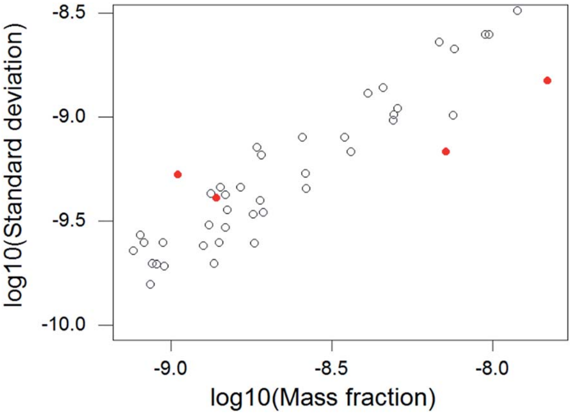

Fig. 1 Reproducibility standard deviations from proficiency tests (circles) and a collaborative trial (red solid circles) in the determination of individual aflatoxins in foodstuffs.

The Horwitz function

In food analysis, however, the shortcomings of individual collaborative trials are offset to a degree by the very large numbers that have been conducted over the years and the generalisations that can be derived from their precision statistics. The Horwitz function is an important generalisation of this type, describing the trend of reproducibility standard deviation (σ_H) as a function of concentration (c), and taking the form

$$\sigma_H = 0.02c^{0.8495},\ 10^{-7} < c < 10^{-1}, \quad (1)$$

with both variables expressed as mass fractions.[5] This function, based on statistics published between the 1930s and 1977, was shown to express the trend of the collaborative trial statistics rather closely at mass fractions between 10^{-7} and 10^{-1} although, of course, it does not predict individual results well because of the scatter around the trend. No amelioration in the trend of precision was discernible in this dataset with the advent of the 'instrumental age' of chemical analysis. Moreover, no substantial improvement in precision was visible in a more recent (1990–2000) collection of statistics.[6] The Horwitz function therefore can be considered as a reasonable summary of collaborative trial statistics for comparisons with those from proficiency tests.

At mass fractions below about 10^{-7} the Horwitz function predicts standard deviations that are inconsistent with detection capability, and in practice we find in that region a tendency for the observed relative reproducibility standard deviation to stabilise at a lower value, centred on 0.22 regardless of concentration.[7] This value is roughly speaking the poorest relative precision that still gives rise to a meaningful result. At mass fractions greater than 10^{-1}, the trend in collaborative trials is again for precisions better than predicted by the Horwitz function.

The data

The statistics used in this comparison are the robust means and standard deviations from all of the qualifying tests provided in the year 2014 by the FAPAS proficiency testing scheme.[8] FAPAS is accredited against ISO/IEC 17043. Only quantities with results expressible as mass fractions were considered. The total number of qualifying tests was 907, encompassing a wide range of analyte types, matrices and mass fractions. The minimum number of laboratories participating in any of these tests was 27 and the median 41. The key to the classification of analytes and matrix types by Series number is shown in the Appendix.

Results and discussion

Mass fractions between 10^{-7} and 10^{-1} (the 'Horwitz region')

Each standard deviation from FAPAS was scaled to (that is, divided by) the value predicted by the Horwitz function for the corresponding concentration of the analyte. A better precision from the proficiency test would result in a scaled value of less than unity. These scaled values are effectively identical to the 'Horrats'[9] used in assessing method performance *via* collaborative trial. The median observed value was 1.01, showing a close relationship between the trends of the precisions of the two sources of interlaboratory information. However, the dispersion of the scaled values was considerable, with a standard deviation of 0.53.

The individual scaled values plotted against mass fraction are shown in Fig. 2. The plot also shows the LOWESS trend of the values. The LOWESS function (locally weighted scatterplot smoother) is a robust model-free trend of the points as a function of mass fraction. There is an overall trend for the scaled statistics to be less than unity at low mass fractions but greater at mass fractions approaching a value of 0.1. This trend is small in relation to the dispersion of the individual values but is significant at 95% confidence by virtue of their large number.

In an attempt to see whether the dispersion of the scaled statistics could be attributed to differences relating to particular analytes or matrix types, they were classified by the FAPAS Series numbers (Fig. 3). FAPAS Series define particular types of analytes, matrices, or methods (see Appendix). There seem to be no

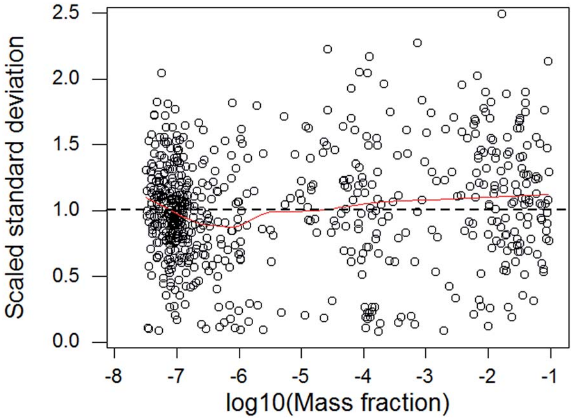

Fig. 2 Scaled reproducibility standard deviation from proficiency tests *versus* mass fraction (points) in the 'Horwitz region', showing the LOWESS trend of the points (solid red line). Eleven high outliers not shown.

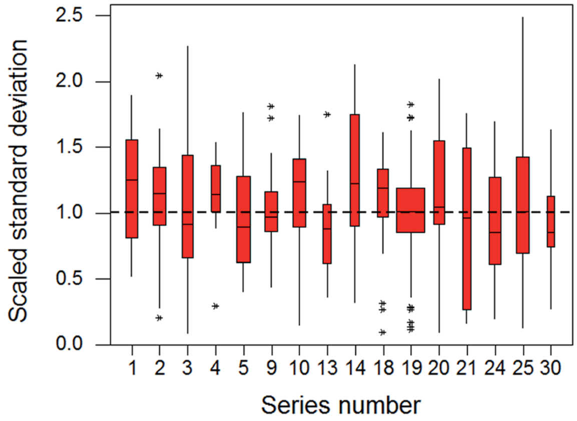

Fig. 3 Scaled reproducibility standard deviation from proficiency tests in the 'Horwitz region', classified according to Series number (key to Series numbers in Appendix). Width of boxes proportional to number of items. Series with less than 10 items omitted.

strikingly discrepant Series, and one-way analysis of variance showed no differences among the Series that was significant at 95% confidence.

Mass fractions smaller than 10^{-7} (the 'low' region)

At mass fractions less than about 10^{-7}, collaborative trials had previously shown[6] a tendency for reproducibility relative standard deviations (RSD_R) to be centred on a value of 0.22, irrespective of mass fraction. This value has been recognised as a suitable modification to the Horwitz function at low mass fractions when used as an analytical fitness for purpose criterion for international trade in food.[10] It was therefore used in this study to scale the RSD_R values from the proficiency test statistics. The outcome is shown in the Fig. 4. The trend of the scaled standard deviations in proficiency tests is here somewhat higher than unity, at an almost constant level, with a median of 1.16 and a standard deviation of 0.55. This corresponds to a median RSD_R of 0.255 in proficiency tests.

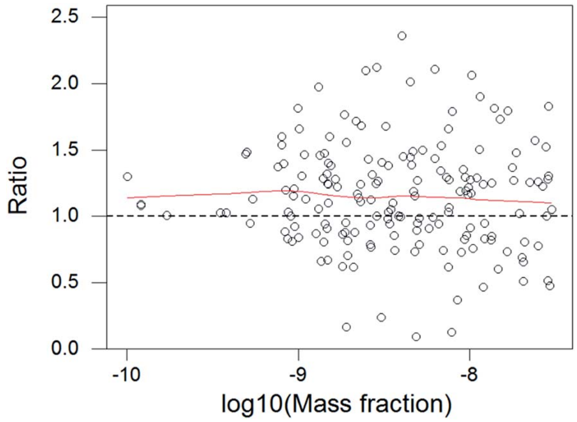

Fig. 4 Scaled reproducibility standard deviation from proficiency tests *versus* mass fraction (points) at 'low' concentrations, showing the LOWESS trend of the points (solid red line). Three high outliers not shown.

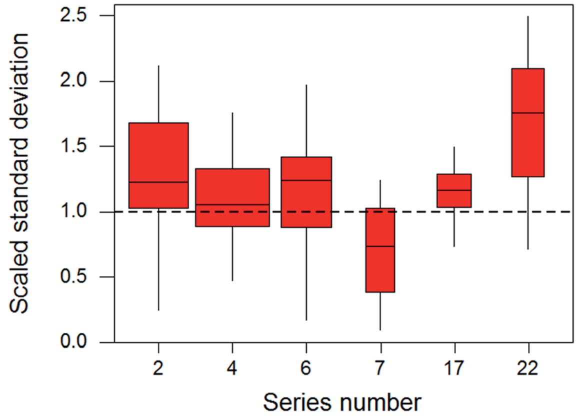

Fig. 5 Scaled reproducibility standard deviation from proficiency tests at 'low' concentrations, classified according to Series number (key to Series numbers in Appendix). Width of boxes proportional to number of items. Series with less than 3 items omitted.

The scaled values were also classified by Series and the outcome is shown in Fig. 5. In this instance there are two apparently discrepant Series, and analysis of variance shows that the variation among the means is significant at 95% confidence. Series 7 involves the determination of trace elements in food, and here proficiency tests provided substantially better precisions than collaborative trials at comparable concentrations. In contrast, Series 22 determinations involve the determination of Fusarium toxins in cereals, and here the RSD_R values from proficiency tests are the greater.

Mass fractions greater than 10^{-1} (the 'high' region)

Fig. 6 shows the scaled reproducibility standard deviations found at mass fractions greater than 10^{-1}. The scaling was executed relative to the predictions of the Horwitz function, regardless of the known tendency of the function to predict values that are too high at these concentrations. As expected, the points show a strong downwards trend as mass fractions

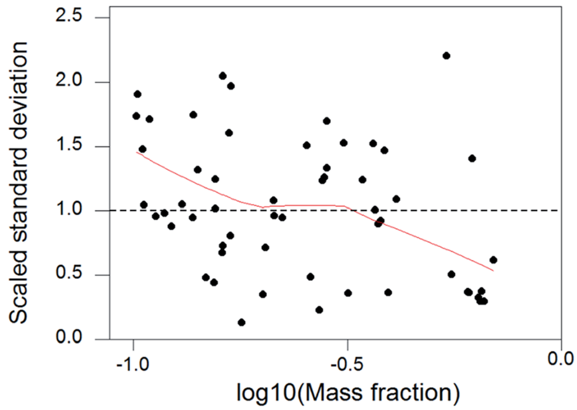

Fig. 6 Scaled reproducibility standard deviation from proficiency tests *versus* mass fraction (points) at 'high' concentrations, showing the LOWESS trend of the points (solid red line). Four high outliers not shown.

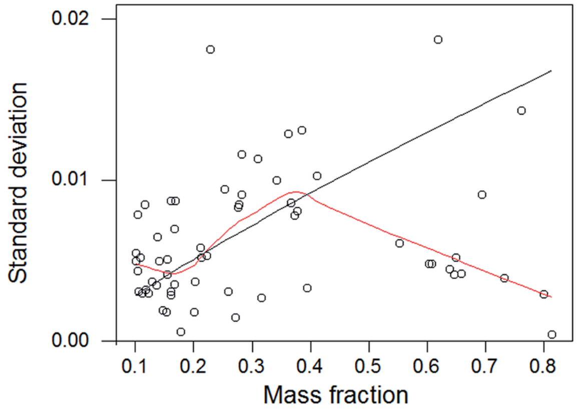

Fig. 7 Reproducibility standard deviation *vs.* mass fraction (points) from proficiency tests at 'high' concentrations, showing the LOWESS trend (red line) and the Horwitz function (black line).

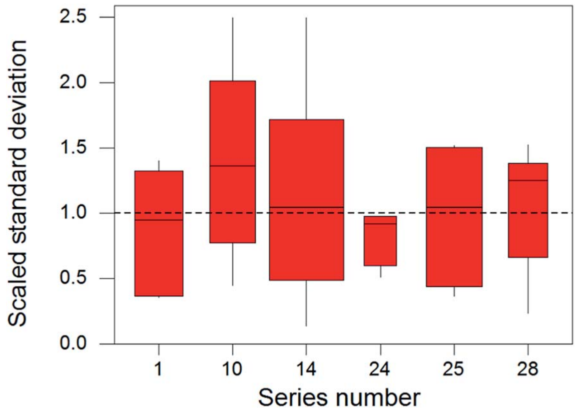

Fig. 8 Scaled reproducibility standard deviation from proficiency tests at 'high' concentrations, classified according to Series number (key to Series numbers in Appendix). Width of boxes proportional to number of items. Series with less than three items omitted.

exceed 0.4. This outcome is perhaps clearer in Fig. 7, where standard deviations are plotted directly. The trend of results follows the Horwitz function well up to a mass fraction of 0.4. The scaled results classified by Series number (Fig. 8) show no visually anomalous classes and analysis of variance shows no variation among the means significant at 95% confidence.

Conclusions

This preliminary comparison between collaborative trials and corresponding proficiency tests has shown that, contrary to expectations, the overall trend is for the two types of interlaboratory study to provide rather similar reproducibility standard deviations at the same concentration of the analyte. However, there is considerable variation among individual values around median levels of the scaled standard deviations, with a standard deviation of about 0.5, showing that case-by-case comparisons (that is, comparisons restricted to specific analyte/matrix combinations) might give rise to a different outcome. Moreover, studies of collaborative trial statistics alone show considerable scatter of individual values around the Horwitz function and its modifications, about double the variation that could be attributed to the random variation inherent in small-number statistics.[5] That alone would inject considerable variation into the scaled values considered in this paper. All of this suggests that a more detailed, case-by-case, comparison would be worthwhile.

Nevertheless, this present study comprises an interesting preliminary upshot showing that, in the food sector, the extra uncertainty in a result brought about by the use of variant procedures or methods is on average relatively small. This fact is becoming increasingly important because the demand for reliable information about the performance of analytical methods is rapidly increasing while at the same time the escalating cost of a collaborative trial is already nearly prohibitive. Proficiency tests, however, thanks to the requirements of accreditation, are becoming ubiquitous and the spin-off information they provide is virtually gratis.

Appendix—key to analytes and matrices by FAPAS Series number

01 Canned meat/meat meal nutritional components.
02 Veterinary drug residues.
03 Soft drinks – components, additives.
04 Aflatoxins and multi-mycotoxins.
05 Pesticide residues, fats and animal products.
06 Polyaromatic hydrocarbons.
07 Metallic trace elements.
09 Pesticide residues, cereals and cereal products.
10 Animal feed, nutritional components and elements.
13 Alcoholic drinks, alcohol content and congeners.
14 Fats and oils, fatty acids.
17 Ochratoxin A, cereals, dried fruit and coffee.
18 Nutritional elements.
19 Pesticide residues, fruit & vegetables.
20 Food additives, permitted and non-permitted.
21 Vitamins, nutritionally important foods.
22 Fusarium toxins and plant toxins.
24 Nutritional components, cereals and cereal products.
25 Nutritional components, milk-based processed foods & fish products.
28 Honey quality parameters.
30 Acrylamide & melamine residues.

References

1 *International vocabulary of basic and general terms in metrology (VIM)*, 3rd edn, JCMG 200, 2008, http://www.bipm.org/vim.
2 M. Thompson, *Anal. Methods*, 2012, **4**, 1598–1611.
3 W. Horwitz, *Pure Appl. Chem.*, 1995, **67**, 331–343.
4 M. Thompson, *Accredit. Qual. Assur.*, 2008, **13**, 479–482.
5 M. Thompson and P. J. Lowthian, *J. AOAC Int.*, 1997, **80**, 676–679.
6 M. Thompson and R. Wood, *Anal. Methods*, 2015, **7**, 377–379.

7 M. Thompson, *Anal. Methods*, 2013, **5**, 4518–4519.

8 FAPAS Secretariat, National Agri-Food Innovation Campus, Fera Science Ltd, Sand Hutton, York, YO41 1LZ.

9 W. Horwitz and R. Albert, *J. AOAC Int.*, 2006, **89**, 1095–1109.

10 *The Codex Alimentarius Commission Procedural Manual*, World Health Organisation/Food and Agriculture Organisation of the United Nations, 24th (English) edn, Rome, 2015, p. 72.