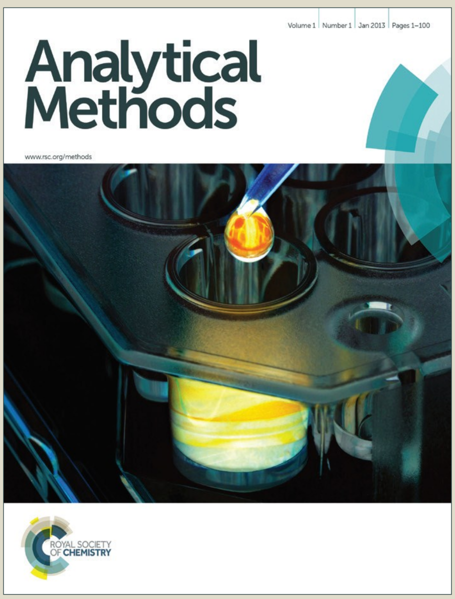
Analytical Methods

Accepted Manuscript



Assessing the stability of a proficiency test material by participant-blind re-use after a period of storage

Mark Sykes
Fera Science Limited
National Agri-Food Innovation Campus
Sand Hutton
York YO41 1LZ
UK

and

Michael Thompson
School of Biological and Chemical Sciences
Birkbeck University of London
Malet Street
London WC1E 7HX
UK

Abstract
A new, powerful and inexpensive test for the stability of reference materials characterised by the results of proficiency tests is demonstrated. Participant-blind re-issue of the same materials provides the scheme organiser with two independent estimates of the location and variance of the results. Because many participant laboratories are involved, the location estimates will not be confounded with changes in the execution of analytical procedures. The method is illustrated by its application to foodstuff test materials almost certain to be stable, with periods between the re-issues of between seven and 52 months. Apart from one discrepant result (out of 41), no statistically significant instability was demonstrated.

Testing the stability of any reference material by direct determination of the analyte after storage under typical conditions is prone to several difficulties. Most obviously, reference materials tend to be specially chosen and prepared to have good stability, so that little change in composition would be evident even after a considerable period of storage. In contrast, the analytical procedures available to observe that change tend to be unhelpfully variable unless a large number of repeat determinations can be made. This means that any economically feasible test for a significant difference between the results before and after storage will have low statistical power, that is, the probability of rejecting a false null hypothesis of stability will be low. A further complication awaits the analyst: all of the analysis should be carried out under randomised repeatability conditions to avoid confusing any changes in the material with changes in the execution of the analytical procedure. Proficiency test providers, however, are in a uniquely powerful position to study the long-term stability of the materials they distribute to participants while avoiding these problems and any undue expenditure.

In the normal course of events proficiency tests compare the results of individual participants, obtained by blind analysis of a test material, with an assigned value that is usually a consensus derived from all of the results. However, they are also able with little effort to obtain a second consensus in a blind re-issue of the same test material after an appropriate interval. Not only is the random variation in a consensus considerably smaller than that of an individual laboratory, but biases of individual laboratories will be largely eliminated in the mean result[1]. Any change with time in the magnitude of an error from a single participant will therefore hardly affect the mean. Even a moderate change in the population of participants would hardly affect the outcome, unless there were a corresponding radical change in the analytical technology that brought about a clear change in performance. It is therefore evident that a statistically significant change in a consensus with time would demonstrate instability. Moreover, the test for instability would have considerably more statistical power than that usually available by virtue of the large number of degrees of freedom.

These ideas are here tested by comparisons of the locations of result sets from a proficiency test produced when the material was issued to participants a second time after various intervals. (A ‘location’ is a measure of central tendency such as a mean, robust mean, median, or mode and is here equivalent to a consensus.)

The data

The data were produced within the FAPAS[2] proficiency testing scheme relating to the analysis of foodstuffs. FAPAS complies with the IUPAC Harmonised Protocol[3] and is accredited against ISO/IEC 17043.[4] Each primary data set comprises results from an average of 68 participant laboratories analysing the same material at two (in one case three) points in time separated by periods between 7 and 52 months. The data were collected with no thought of stability testing at the time. The materials were completely re-labelled between rounds so there was no possibility of the results being affected by previous knowledge accrued by the participants. Indeed participants are warned in a general way not to take account of previous results because closely similar but different materials may be issued without warning to deter just such malpractice.

The statistics considered here are the robust means and standard deviations of the valid results submitted by the participants. Results are valid unless they are or seem to be submitted in the wrong units or are otherwise wildly far from the consensus. Robust methods are used to reduce the influence in the statistics of any remaining outlying results and the heavy tails that are not rare in proficiency test data. The robust procedure used was Huber’s proposition H15[5]. These statistics are produced as routine in FAPAS (not especially for the present study) and are as listed in the original reports.

The analytical measurements considered here are determinations of proximate analytes, (moisture, ash, fat, and nitrogen) in foodstuffs resembling meat pies in average composition. In some instances other proximate analytes were also determined. The reason for this selection is that these were almost the only instances where the test materials were issued twice. This recurrence stemmed from economic considerations: the preparation and canning of this type of test material is costly to set up and therefore best conducted in large batches that are then available for distribution in two or more rounds of the Series. While the availability of these test materials for the present purpose was serendipitous, they could hardly have been better chosen for the demonstration: they were cooked and hermetically sealed by canning, so were insusceptible to bacterial decomposition and photolysis. As such they were *a priori* almost certain to be stable and therefore ideal to test the stability test itself.

Results and discussion.

A pair of primary datasets is shown in Fig. 1. The identities, and robust means and standard deviations of each dataset are shown in Table 1. A test statistic was calculated for each pair of datasets as follows: $z^* = (\hat{\mu}_1 - \hat{\mu}_2)\big/\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$ where $\hat{\mu}_1, \hat{\mu}_2$ are the robust means of the earlier and later datasets respectively, and $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ are the corresponding robust variances. If there were no bias between the two means, the outcome would be a z^* approximately following the standard normal deviate, that is, with a mean of zero and a variance of unity. The z^* statistics found are shown in Table 1. (*Note:* z^* must not be confused with proficiency test z-scores.)

A value of z^* outside the range ± 2.5 suggests an individually-significant outcome. Considering the results as a whole, it is clear that there is one discrepant value of $z^* = 4.1$. There is no obvious explanation of the discrepancy which, in any event is not unduly large (a change of -1.3 % relative in the concentration of nitrogen. The discrepant datasets are shown in Fig 1. The remaining 40 test statistics, having a mean of 0.02 and a standard deviation of 1.17, are by the Kolmogorov-Smirov one-sample test consistent with a normal distribution and an overall null hypothesis of stability. The results show no trends attributable to the length of storage or the identity of the analyte (Fig 2).

Conclusions

A method of testing for the stability of a reference material characterised by the robust mean and variance derived from two rounds of a proficiency test has been demonstrated. The test materials were particularly apposite for this demonstration because they were *a priori* almost certain to be stable: they were cooked and hermetically sealed by canning, so were not susceptible to bacterial decomposition or photolysis. In the event, apart from possibly one instance, the results (individually or combined) showed no significant deviation from stability, demonstrating the power and accuracy of the test. In routine applications of the method in general, of course, there would be no prior strong expectation of stability.

Table 1. Means and standard deviations from pairs of rounds of a proficiency test in which the same test material was distributed, showing the interval between the rounds and the resulting test statistic. Results are expressed in mass fractions as a percentage. ID_1 and ID_2 are respective FAPAS Round identification numbers.

Analyte	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	ID_1	ID_2	Interval (months)	$(\hat{\mu}_1 - \hat{\mu}_2)$	Test statistic z*
ash	3.140	0.010	3.171	0.005	158	173	32	-0.031	-2.766
ash	2.848	0.006	2.832	0.007	172	175	7	0.016	1.743
ash	1.397	0.006	1.400	0.004	156	177	45	-0.003	-0.462
ash	2.940	0.007	2.930	0.008	164	181	35	0.010	0.911
ash	1.268	0.006	1.280	0.006	154	183	62	-0.012	-1.416
ash	2.060	0.006	2.070	0.009	178	184	12	-0.010	-0.912
ash	2.066	0.006	2.060	0.006	160	178	39	0.006	0.698
ash	2.778	0.006	2.770	0.007	161	186	52	0.008	0.881
chloride	0.600	0.004	0.599	0.003	156	177	45	0.001	0.132
chloride	0.416	0.007	0.414	0.004	154	183	62	0.002	0.302
chloride	0.744	0.009	0.772	0.010	178	184	12	-0.028	-2.161
chloride	1.342	0.007	1.320	0.012	161	186	52	0.022	1.574
hydroxyproline	0.812	0.011	0.813	0.013	164	181	35	-0.001	-0.082
moisture	65.419	0.045	65.468	0.046	158	173	32	-0.049	-0.758
moisture	70.565	0.037	70.642	0.033	172	175	7	-0.077	-1.539
moisture	80.846	0.036	80.800	0.028	156	177	45	0.046	1.007
moisture	59.781	0.053	59.620	0.067	164	181	35	0.161	1.881
moisture	76.437	0.035	76.400	0.033	154	183	62	0.037	0.770
moisture	60.030	0.059	59.900	0.086	178	184	12	0.130	1.247
moisture	60.071	0.047	60.030	0.059	160	178	39	0.041	0.543
moisture	71.994	0.068	72.000	0.094	161	186	52	-0.006	-0.051
nitrogen	3.652	0.008	3.667	0.007	158	173	32	-0.015	-1.454
nitrogen	2.454	0.006	2.421	0.006	172	175	7	0.033	**4.124**
nitrogen	1.201	0.004	1.210	0.003	156	177	45	-0.009	-1.717
nitrogen	3.498	0.008	3.510	0.009	164	181	35	-0.012	-1.036
nitrogen	1.542	0.004	1.540	0.004	154	183	62	0.002	0.355
nitrogen	2.260	0.007	2.270	0.010	178	184	12	-0.010	-0.794
nitrogen	2.264	0.006	2.260	0.007	160	178	39	0.004	0.434
nitrogen	1.560	0.006	1.560	0.007	161	186	52	0.000	0.021
sodium	0.377	0.004	0.377	0.003	156	177	45	-0.000	-0.084
sodium	0.271	0.004	0.267	0.002	154	183	62	0.004	0.796
sodium	0.592	0.007	0.602	0.008	178	184	12	-0.010	-0.945
sodium	0.871	0.009	0.867	0.008	161	186	52	0.004	0.364
total fat	9.071	0.050	9.157	0.056	158	173	32	-0.086	-1.139
total fat	8.958	0.042	8.838	0.047	172	175	7	0.120	1.888
total fat	0.805	0.016	0.760	0.016	156	177	45	0.045	1.999
total fat	16.102	0.102	15.960	0.111	164	181	35	0.142	0.940
total fat	6.612	0.059	6.550	0.041	154	183	62	0.062	0.856
total fat	17.230	0.096	17.500	0.106	178	184	12	-0.270	-1.894
total fat	17.238	0.087	17.230	0.096	160	178	39	0.008	0.065
total fat	11.423	0.093	11.400	0.109	161	186	52	0.023	0.164

[1] M Thompson, L Owen, K Wilkinson, R Wood and A Damant. *Analyst*, 2002, **127**, 1666-1668
[2] FAPAS Secretariat, Fera Science Limited, National Agri-Food Innovation Campus, Sand Hutton,York, YO41 1LZ, UK.
[3] M. Thompson, S. Ellison and R. Wood, *Pure Appl. Chem.*, 2006, **78**, 145
[4] ISO/IEC 17043:2010, Conformity assessment – General requirements for proficiency testing
[5] Analytical Methods Committee. *Analyst*, 1989, **114**, 1693-1697.

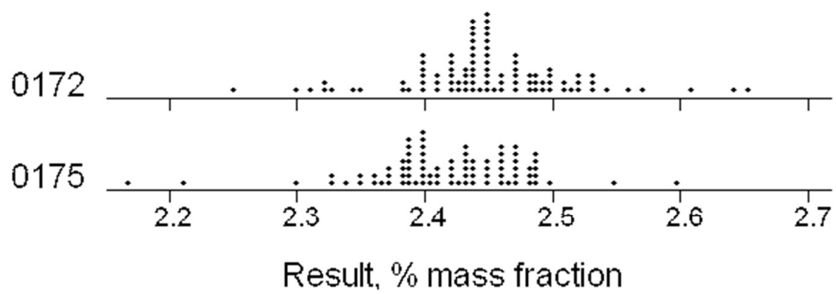


Fig 1. Results of participants for nitrogen in rounds 0172 and 0175.

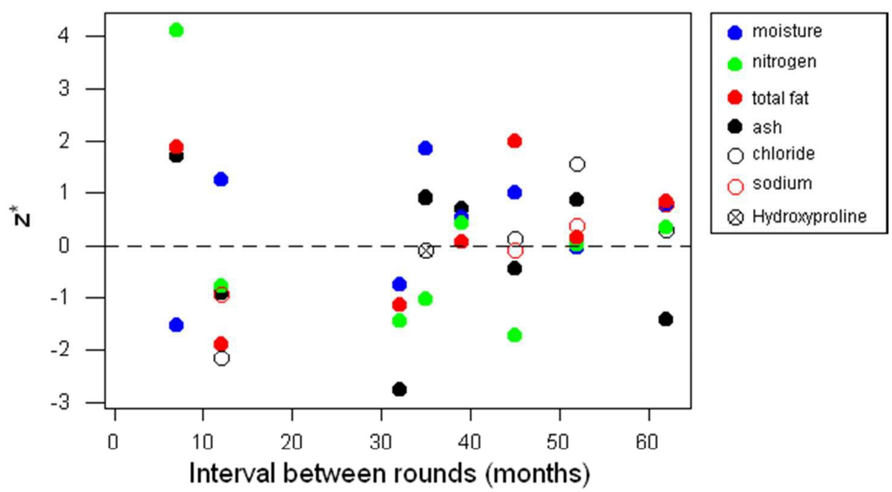
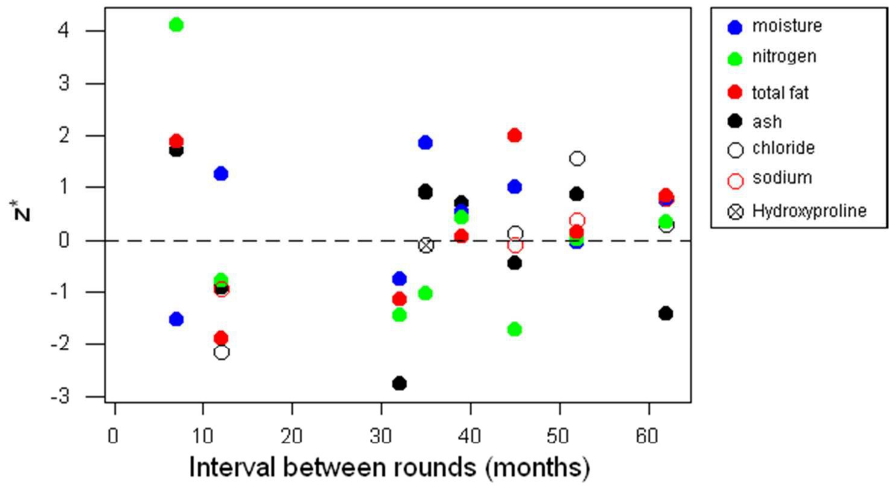


Fig 2. Test statistics *z** (points) for proximate analytes by time of storage and identity of the analyte. A value of *z** outside the range ± 2.5 suggests an individually-significant outcome.

Graphic and text for graphical abstract.



Stability of a reference material characterised by proficiency test is monitored inexpensively by a participant-blind issue of the same material in two rounds of the test separated by a suitable period.