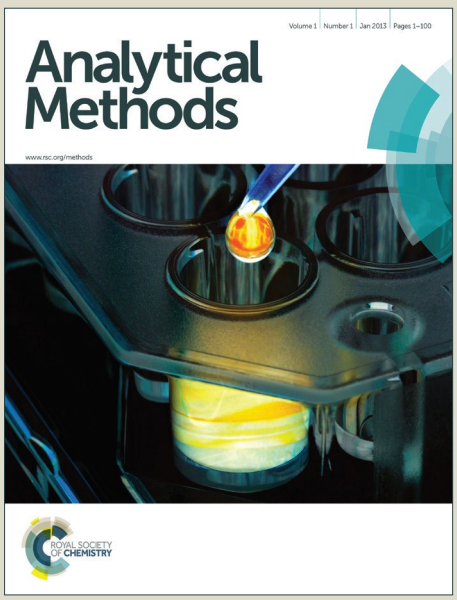
Analytical Methods

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Is your 'homogeneity test' really useful?

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

Abstract

'Homogeneity testing' is a formal requirement in the preparation of reference materials for certification and for proficiency testing. Few scientists, however, seem to be aware of the rather severe limitations that apply to the outcome of any such test of a size that is economically feasible. Typically the tests have low statistical power to detect significance, and the resulting estimates of between-bottle standard deviation have wide confidence limits. Scientists should bear these limitations in mind and avoid being over-prescriptive when drafting standards and guides.

Reference materials are essential in chemical measurement for demonstrating comparability of results over space and time. Such materials are usually divided into many portions (here called 'bottles' to conform to ISO Guide 35[1]), in activities such as proficiency testing, collaborative trials, internal quality control and the preparation of certified reference materials. Clearly all bottles of a reference material should have the same composition within margins that are optimally narrow in relation to the mass of a test portion and the cost of using the material. Some degree of testing is needed to show how closely that need is fulfilled. The test naturally adapted to the task is based on experiments with randomised replicated measurements, followed by analysis of variance (ANOVA).

ISO Guide 35 details the measures required to ensure that the test provides an unambiguous outcome. For instance, a random selection of bottles is essential to make the expectation of the test representative. A further requirement is the need to ensure that compositional variation among bottles of the test material is not confounded with systematic variation (drifts or saltations, for example) in the measurement procedure. For this reason the Guide recommends a random order in the sequence of measurements on the test portions or tests for time trends in the outcome. However, an unexpected consequence of the Guide is that it has encouraged the unwary to believe that its recommendations are sufficient as well as necessary. There is a tendency for scientists to believe that working according to the prescriptions of the Guide frees the user from the further responsibility of looking at all of the implications and examining the features that determine whether a particular design of homogeneity test is worthwhile.

This paper is designed to redress the balance. First let's dwell on some 'home-truths' about homogeneity and ANOVA that impinge on our discussion. If we ignore these realities we become prone to misconceptions about why we should conduct homogeneity tests and how we should apply the results.

- All reference materials are heterogeneous, at least in principle and, nearly always, in practice. The important criterion for analytical chemists is that the inevitable variation in composition between bottles, determined on appropriately-sized test portions, is sufficiently small in relation to an uncertainty consistent with fitness for purpose. This has given rise to the apt but clumsy designation 'sufficiently [close to] homogeneous' in preference to 'homogeneous'.
- There is no test for homogeneity: this is despite the almost universal misuse of the phrase 'homogeneity testing'. We *can* test for a statistically significant degree of heterogeneity, which is not the same thing, but we must reconcile ourselves to the fact that an affordable test is unlikely to be sufficiently powerful.
- The tests also provide estimates of between-bottle dispersion that (a) are likely to differ wildly from the true value and (b) are very sensitive to mildly unusual results unless robust ANOVA is used.

Significance, importance, and power

Tests for significant heterogeneity in chemical measurement that are on an economically-feasible scale are notoriously low in statistical power[2]. This means in practice that the test is unlikely to detect consequential degrees of heterogeneity at 95% confidence in a large proportion, perhaps a substantial majority of instances. The bigger the experiment, and the more precise the analytical results, the better it will be for characterising an important degree of heterogeneity, but the more it will cost to conduct.

The overall message that emerges is that it is usually easy to render a material sufficiently close to homogeneous, but difficult to find an analytical procedure with sufficient precision adequately to explore the real variation between bottles by means of an experiment of economically-feasible size. This is perhaps made clearer when the outcome of the test is focussed on the magnitudes of the components of dispersion rather than on statistical significance. We find that the estimates of the variance components other than analytical are unreliable in magnitude. For instance, proficiency test providers often use a test with duplicate analysis on 10 bottles of the material, a (10×2) test. Simulations can show us what to expect under the assumption that the analytical results were independent, random, normally-distributed variables with repeatability standard deviation σ_r. By conducting such a test on a hypothetical material that was indeed homogeneous (that is, the true between-bottle standard deviation $\sigma_b = 0$), we find that the 95% confidence limits on the estimated value $\hat{\sigma}_b$ would be $(0, 0.82\sigma_r)$, . About half of the estimates would be zero, but the mean would be about $0.24\sigma_r$ and results as high as $0.82\sigma_r$ would not be rare. The outcome is hardly better when the between-bottle standard deviation is as large as $\sigma_r/2$.

Information on between-bottle variance by using enhanced power—a unique example

This shortcoming has recently been demonstrated in a study of heterogeneity in foodstuff proficiency-test materials, namely products similar in composition to meat pies[3]. Two analytes were selected for this pilot study, nitrogen (a proxy for protein) and fat. Nitrogen was considered because its concentration in successive test materials was restricted to a small range, so each variance component could reasonably be taken as homoscedastic (that is, of uniform variance). The mass fraction of fat, in contrast, varied between about 2% and 30%, but it was found that the variances could be rendered close to homoscedastic by considering standard deviations (SD) scaled to concentration (that is, as relative standard deviations (RSD)). (Homoscedasticity is required to fulfil the requirements of ANOVA model.)

Data from routine testing (10×2 experiments) in 20 successive rounds of the proficiency test were analysed by hierarchical ANOVA. This boosted the degrees of freedom for the between-bottle variance 20-fold, greatly increasing the power of the statistical analysis and providing for the first time plausible estimates of the corresponding standard deviations. For nitrogen the between-bottle SD estimate was 0.008% mass fraction (about one third of the analytical SD). For fat the between-bottle RSD was 0.006, again about one third of the corresponding analytical RSD.

Bodies that organise the preparation of reference materials go to great lengths to ensure that the resulting material is as close to homogeneous as can be reasonably achieved. As shown here, with a carefully prepared material and with ten bottles analysed in duplicate, the routine test is usually unable to detect the low level of heterogeneity. Moreover, the estimate of between-bottle variance is unreliable, being inordinately disperse, and often substantially biased.

A larger example, taken from ISO Guide 35, Appendix B3

The example depicts results of a test for between-bottle heterogeneity of chromium in a certification study. The study is of substantial size (20×3), that is, requiring the analysis of 60 separate test portions. The data are shown graphically in Fig 1 and show a variation between bottles that is significant at 95% confidence. Calculations displayed in the Guide show that $\hat{\sigma}_b = 3.93\,\text{mg kg}^{-1}$. However, even with this large experiment and a statistically-significant outcome, the estimate $\hat{\sigma}_b$ is very variable, with 95% confidence limits, estimated by the bootstrap[4], of (1.78, 5.33).

The outcome is also sensitive to the presence of mild outliers. For example, if just one of the 60 data is changed to a result of 140 mg/kg, as shown in Fig 1, the result of the ANOVA becomes not significant at 95% confidence and the estimate is considerably changed, to $\hat{\sigma}_b = 1.82$ with 95% confidence limits of (0.00, 4.05) mg/kg. The influence of a small proportion of analytical outliers can be largely eliminated by the use of a robust ANOVA. In these applications, however, it is essential to ensure that the robustification is applied only at the within-bottle level of the design: it then accommodates only analytical outliers. Between-bottle variation represents real heterogeneity and variation at that level must not be robustified. Application of such an ANOVA to the modified data above gives estimates that are close to those of the original data, namely $\hat{\sigma}_{b,robust)} = 3.85$ with 95% confidence limits of (2.34, 5.56) mg/kg.

Heterogeneity and certification of a reference material

The 'official' procedure for attaching an uncertainty to a certified value for a reference material includes a term for between-bottle (and sometimes within-bottle) heterogeneity[1]. This is formally correct if we know the population parameter. But we have only an estimate of the parameter and, as we have seen just above, the estimate is likely to be seriously in error. Using such a term for comparison with a separate fitness-for-purpose criterion could be badly misleading. When the overall uncertainty on the certified value is estimated, the difficulty can be obviated by subsuming the heterogeneity contribution into the repeatability or reproducibility standard deviation.

We should also consider whether it is worthwhile separately to study the within-bottle heterogeneity. Producers of certified reference materials are encouraged to do this presumably on the grounds that it contributes to repeatability dispersion. But the estimates of this statistic are likely to be at least as unreliable as between-bottle estimates and inflate the cost of the test twofold. Wouldn't the same financial layout be better used by doubling the number of bottles in the experiment and ignoring within-bottle variation? There is no general answer to this question. We must further remember that many 'homogenised' materials have a tendency to segregate in their containers during transport and use. Therefore even a hypothetically-reliable measure of within-bottle heterogeneity carried out by the producer will be untrustworthy by the time the bottle is on the analyst's bench. We should also bear in mind that, in routine practice, it is part of the analyst's job to ensure as far as possible that the test portion taken is representative of the laboratory sample (using both words as strictly defined[5]). This responsibility should apply equally to use of reference materials. Now risk of contamination obviously prevents an analyst from removing a certified reference material from its container for further treatment such a grinding. A thorough shaking within the closed vessel is the most that could be sanctioned, and is always an essential preliminary before removal of the test portion. It would be clearly useful to know

that a preliminary shaking might restore the bottle's contents to sufficient homogeneity, but the cost of that knowledge is likely to remain prohibitively large.

Finally we should consider whether in testing for heterogeneity it is worthwhile to use more elaborate experimental designs, such as conducting the analysis in randomised blocks. This measure can be useful when results are prone to suffer from drift in long runs of analysis or when a sufficiently long sequence of analysis cannot be accommodated in a single run. (A 'run' is the period during which repeatability conditions can be assumed to prevail.) This would tend effectively to improve the precision of the analytical results. A minor disadvantage of a blocked design would be that a few degrees of freedom are wasted on a feature that does not otherwise affect the question at issue, the heterogeneity of the candidate material. Unfortunately including a run-to-run effect is apparently often misinterpreted as an inherently desirable feature of the test for heterogeneity, whereas each instance should be considered on its merits.

Preliminary tests of datasets for deviations from procedure

It should be regarded as essential practice to test datasets from heterogeneity studies to ensure that they show no obvious sign that the analyst had deviated from the selected design so as to invalidate the outcome of the ANOVA. Deviations from a strictly random ordering of the test portions in an analytical procedural sequence are not uncommon. Features often encountered by the author include (a) a trend or step-change in the results ordered by the bottle numbers or by order of analysis, (b) a systematic difference between the first and second results on each bottle, (c) insufficient digit resolution in the data for a reliable statistical analysis, and (d) outlying differences between duplicate results. These features tend to emerge when the instructions for conducting the experiment are insufficiently detailed. Many of these problems can be immediately spotted on a graph of the data, organised in the order of analysis and (as in Fig 1,) organised by bottle number, and confirmed by a test of significance.

Conclusions

'Homogeneity tests' usually fail to deliver what at first sight they seem to promise. Even with datasets faultlessly produced according to an appropriate experimental design, the tests seldom detect significant heterogeneity because experiments that are economically feasible have insufficient power. They deliver estimates of between-bottle standard deviation that are wildly variable, biased, and very sensitive to outlying analytical results. Awareness of these problems would be reinforced if the statistical packages delivering ANOVA provided confidence limits on the estimated standard deviations . ANOVA, robustified at the analytical level but giving these confidence limits, would be a invaluable tool for 'homogeneity testing'. That would alert analysts to the possibility of improper interpretation of the outcome.

A test for heterogeneity provides important reassurance to users and as such can hardly be dispensed with. But it is suited better to screening materials for complete failures in the homogenisation process or other type of mistake than to studying heterogeneity *per se*. In proficiency tests, where 10×2 tests are at the limit of affordability, there is no question of doing more. In the preparation of certified reference materials, a considerably greater expenditure may be justifiable, but even here the outcome may fall short of providing all of the information that analysts would like to see.

There is an unfortunate relationship between the heterogeneity of a reference material and the precision of the analytical procedures that it is designed to control. To fulfil its purpose adequately, the material must have a between-bottle dispersion that is small in relation to the analytical variation of the procedure that it is used to monitor. But in that circumstance, that particular analytical procedure will be unable to detect the heterogeneity in a reasonably-sized experiment. In many, perhaps most instances, there will be no economically acceptable alternative analytical procedure with better precision. For between-bottle heterogeneity in a reference material, 'if you can measure it properly, it's too big'.

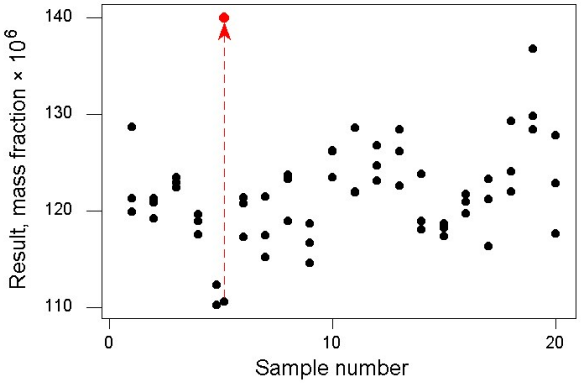


Fig 1. Results (black solid circles) for the triplicate analysis for chromium of 20 bottles of a candidate reference material. The red solid circle indicates a modification of the data to investigate the influence of a single result on the estimate of between-bottle standard deviation.

[1] ISO Guide 35: 2006. *Reference materials -- General and statistical principles for certification.*

[2] M Thompson. *Accred Qual Assur*, 2008, **13**, 581-584.
[3] M Thompson and T Fearn. *Analytical Methods*, 2011, **3**, 2529-2533
[4] Analytical Methods Committee. *AMC Technical Briefs* No. 8 (2001). *The bootstrap: a simple approach to estimating standard errors and confidence intervals when theory fails.* (Download gratis via www.rsc.org/amc)
[5] M H Ramsey and S L R Ellison (eds), Eurachem/Eurolab/CITAC/ Nordtest/AMC Guide. *Measurement uncertainty arising from sampling. A guide to methods and approaches.* Eurachem 2007.