Comparing Uncertainties – Are they really different?[1]

Peter D. Rostron
School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, UK
Tel: 07956 765 205
email: pr52@outlook.com

Tom Fearn
Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK

Michael H. Ramsey
School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, UK

Abstract

Uncertainties occur at all stages of a measurement process. Quantification of these uncertainties is important in order to make reliable decisions based on these measurement results. In some cases it can be useful to be able to compare the uncertainties associated with different measurement methods, in order to establish the method that is most reliable. A comparison can also be made between uncertainties that have themselves been evaluated using different estimation procedures. This paper discusses the comparison of uncertainties in chemical measurements using case study examples. Depending on the context, both exact and approximate *F*-tests are used to compare the ratios of uncertainties, while in some cases the approach is to compare separate confidence intervals.

Keywords
Measurement uncertainty, duplicate method, robust ANOVA, confidence interval

Introduction

It has become accepted practice for laboratories to report an uncertainty within each measurement result (i.e. measurement uncertainty, MU). This is an indication of the quality of the results obtained in chemical analysis, for example, and is important in the evaluation of the fitness for purpose of the results obtained by a particular measurement process. Examples of where knowledge of MU is important are for internal quality control purposes (e.g. analytical repeatability), evaluating regulatory compliance and conformity, and assessment of between-lab reproducibility [1, 2]. Measurement uncertainty is formally defined as a 'parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand' [3], where the measurand is defined as the 'Quantity intended to be measured' [4]. The value of the measurand is equivalent to the true value of the analyte concentration.

In many cases in chemistry, it is the analyte concentration in the sampling target (the portion of material, at a particular time, that the sample is intended to represent [5]) that is of interest. It would be less common that the determination of the analyte concentration in the laboratory sample is the primary objective. Uncertainties can occur at all stages of a measurement process, including the initial taking of a primary sample from a sampling target. There is increasing awareness that uncertainty in the sampling process is often the dominant component of the measurement uncertainty. Sampling uncertainty encompasses the uncertainty in the sampling process, transportation, and also in the sample preparation before it is subject to chemical treatment in the laboratory [5]. It is therefore useful to be able to separate the measurement uncertainty into its two encompassing components: uncertainty due to sampling, and uncertainty due to chemical analysis, where the latter is sometimes estimated as analytical repeatability or reproducibility.

[1] Based upon a presentation by the first author at the EURACHEM Workshop 'Uncertainty from sampling and analysis for accredited laboratories', November 2019, Berlin, Germany

When measurements of the analyte concentrations in a number of different sampling targets are made, (e.g. within a bulk of material), the *duplicate method* can be used to empirically quantify the uncertainties due to sampling and analysis [5]. In this procedure a number of the sampling targets are selected at random for duplicate sampling, and each sample (including duplicates) is subject to duplicate analysis (Fig. 1). Duplicates are recommended at a minimum of 8 sampling targets [6]. The sampling and analytical components of the overall measurement uncertainty can then be evaluated separately using nested analysis of variance (ANOVA).

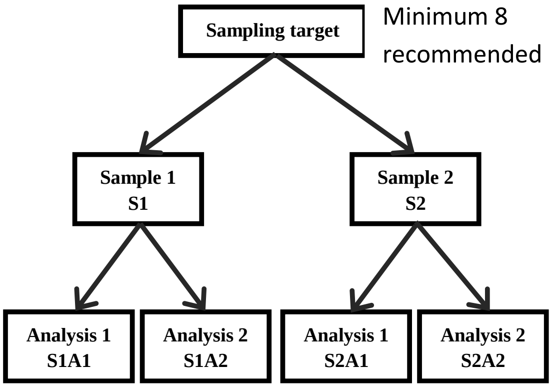

Fig. 1 Balanced design used for evaluating sampling and analytical components of uncertainty in the duplicate method [5]

An alternative to the traditional, or classical ANOVA, is the use of robust statistics. It is often the case that a set of measurements contains a small proportion of outlying values, which can have a disproportionate effect on the mean and standard deviations calculated by ANOVA. In this case the use of robust ANOVA is recommended in the Eurachem guide [5]. Robust ANOVA is an iterative process that does not remove, but down-weights the effects of a small proportion (< 10 %) of outlying values. It can only realistically be achieved using a computerised technique [7]. A computer program called RANOVA3 is available on the website of the Analytical Methods Committee of the Royal Society of Chemistry (UK) that performs both classical and robust ANOVA on the type of 3-tier nested experimental design shown in Fig. 1 [8]. One potential application of this dual capability is that a large difference between classical and robust variances indicates the presence of outliers in the data. The magnitudes of both sampling and analytical uncertainties depend on the particular sampling and analytical methods used. In certain situations it may be useful to be able to compare uncertainties that have been evaluated by different sampling or analytical methods.

For example, reference materials that are used to evaluate the bias of an analytical method are usually supplied with data sheets that specify the uncertainties of the quoted (sometimes certified) concentrations of key analytes in the material. Such quoted uncertainties are typically based on an assumption of a minimum mass of material that will be analysed using a laboratory based method. There is an increasing use of technology developments such as Portable X-Ray Fluorescence (PXRF, which can also analyse *in situ*) and laser ablation methods, both of which analyse a very small test-portion mass for each measurement. Consequently, the measurement uncertainty may then be found to be larger than the quoted value that typically applies to more conventional laboratory methods. This can arise because heterogeneity of the analyte may occur at a similar, or larger, scale than the mass or spatial extent of a very small test portion in the milligram or microgram scale. The questions then arise whether the uncertainty evaluations are significantly different between the different analytical methods, and also whether the measured value is significantly different from the quoted value (See Case Study 1). Another example where it might be useful to compare uncertainties is in the assessment of new or different methods of uncertainty evaluation (see Case Study 3).

Testing the equality of two variances is straightforward when their estimates are the sample variances of two independent normal samples. Then an *F*-test of the variance ratio is valid. When the normality assumption does not hold because of the presence of outliers it may still be reasonable to use the *F*-test on the ratio of two robustly estimated variances, though this test will now be approximate. The problem becomes more difficult when the variances being compared have been estimated from one or more experiments like the nested design shown in Fig. 1, either as individual components of variance at any level above the bottom level (analysis level in Fig. 1), or as linear combinations of these components of variance. The issue here is that the ratios of variance estimates will only have an *F*-distribution under the null hypothesis of equality when the estimates themselves are independent, and each is distributed as a multiple of chi-squared. These two conditions require normal distributions for the variability at all the relevant levels of the data, but even then only apply in some special cases, essentially when each variance estimate is derived from a (different) single mean square in the natural ANOVA(s) of the experimental data. Some of these cases are discussed in the examples below. In situations where the *F*-test is not strictly correct but where the variance component of interest is much larger than the ones below it in the hierarchy it may be possible to use an approximate *F*-test, but for the general case we need another approach. One option would be to compute a bootstrap distribution for the variance ratio, either for a classical analysis or for a robust analysis. This option is not currently available in RANOVA3, and the bootstrapping approach is only reliable given a sufficient sample size (see discussion below), but is a possibility for the future.

Another option is to construct, say, 95 % confidence intervals for each of the variances to be compared and see whether these intervals overlap. If they do not, then the variances are different with a significance level of $p < 0.05$. In the cases where the *F*-test is incorrect these confidence intervals will necessarily be approximate, but they are good approximations when the number of sampling targets (*n*) is large. For example, simulations with $n = 100$ and a nominal coverage percentage of 95 % gave average coverage percentages ranging between 94.5 % – 95.4 % for the 3 levels in Fig, 1. Reducing the number of sampling targets to $n = 10$ resulted in smaller actual coverage percentages, calculated by simulation as 89.0 % – 92.4 % [9]. A more serious problem is that the comparison of confidence intervals has low statistical power as a test [10]. For example, comparing two variances from separate normal samples of size 25 using the *F*-test needs an observed variance ratio of 2 to achieve significance at $p = 0.05$, whilst non-overlap of the separate 95 % confidence intervals would need an observed variance ratio that is greater than the ratio of the upper and lower points on a chi-squared distribution with 24 degrees of freedom, i.e. 39.36/12.40 = 3.17. This is not a problem if the confidence intervals are either non-overlapping or massively overlapping, but a small overlap is likely to leave the analyst wishing for a more powerful test. In the examples that follow both exact and approximate *F*-tests, as well as the comparison of confidence intervals, are illustrated.

The objectives of this paper are as follows:

Objectives

1. Describe the calculation of confidence intervals (CIs) for uncertainties both for normal and non-normal data.
2. Demonstrate the comparison of two uncertainty evaluations for normally distributed data, using both a formulaic approach (*F*-test) and also comparison of CIs (Case study 1).
3. Demonstrate the comparison of two uncertainty evaluations for non-normally distributed data using comparison of CIs calculated using a bootstrapping procedure (Case studies 2 and 3).
4. Discuss the strengths and weaknesses of these methods.

Calculation of confidence intervals of uncertainties

Mathematical approaches for determining CIs of variances from ANOVA exist, however these are based on probability models that assume the data arise, at least approximately, from normal distributions. When the data are significantly non-normal, these models for determining confidence intervals are not appropriate, and a different approach is needed. One solution is to use a bootstrapping procedure. In brief, a large number of independent bootstrap samples (e.g. 2000) of the same size and structure as the observed dataset are generated by random sampling with replacement. For the design in Fig. 1, each bootstrap sample is generated as follows:

1. For *n* sampling targets, *n* means at the sampling target level are selected at random with replacement.
2. For each of these *n* mean values, two differences at the sample level are selected at random with replacement, and used to generated 2 new values at the sample level, giving a total of 2*n* new sample values.
3. For each of these values at the sample level, two differences at the analytical level are selected at random with replacement, and used to generate 2 new values at the analytical level, giving a total of 4*n* new analytical values.

This procedure results in a bootstrap sample of the same structure as the original sample, i.e. 4 simulated analytical measurements for each of *n* sampling targets. The statistic of interest is calculated for each bootstrap sample and stored. The resultant values are sorted into numerical order. Confidence intervals can then be derived from the empirical distribution of the sorted values. This approach has been developed to calculate the CIs for the variances for the robust algorithm used in the RANOVA3 program, for a balanced 3-tier nested design such as shown in Fig. 1. It has been shown to work best in situations where the number of sampling targets is large or the number of outlying values is small [9]. As previously stated, non-overlap of CIs for 2 different variances indicates that the 2 variances are significantly different.

As noted above, this type of test has low power. If the CIs overlap it may be that a more powerful test would have shown a significant difference. A potential future advancement to this approach would be to develop bootstrapping routines to give upper and lower confidence limits for ratios of variances. If this CI for a particular variance ratio contains the value 1 then there would be insufficient evidence to conclude that a difference exists between the two variances.

Case Study 1: Comparison of heterogeneity in reference materials measured using PXRF at two different beam sizes (normally distributed data)

Heterogeneity of the analyte concentration is one of the most important factors contributing to the uncertainty in sampling. It is defined as ‘The degree to which a property or constituent is uniformly distributed throughout a quantity of material’ [11]. Different samples taken from different locations within a sampling target that has a heterogeneous analyte distribution are likely to have varying concentrations of analytes, thereby increasing the contribution of sampling to the overall measurement uncertainty [5]. The following case study has been drawn from a study intended to quantify the uncertainty due to heterogeneity in small quantities of a reference material [12]. It was found that the degree of heterogeneity tended to increase with decreasing sample mass.

Reference materials (RMs) are used to evaluate the bias of measurements made by an analytical method, often as part of equipment calibration and also where measurement traceability is required. A series of three blended RMs (known as the SdAR series [13]) has been designed to resemble soils or sediments with different levels of environmental contamination from mining operations. Each reference material is accompanied by a data sheet that states the reference value of each element or compound for which an analyte concentration value is assigned, with an associated uncertainty. It is further stated that the minimum recommended test portion size is 0.2 g. This number is based on the results of homogeneity tests and also assessments of the methods used in a proficiency testing scheme [13].

The minimum test portion mass is considered practical for most laboratory analytical methods. However, there is an increasing use both of field portable equipment such as PXRF devices, and also laboratory technologies such as secondary-ion mass spectrometry (SIMS) methods that use very small test portions in the picogram range [14]. The former enable a much more rapid turnaround in obtaining measurement results, require minimal sample preparation, and in many cases can be used to make measurements *in situ*.

One potential disadvantage of these methods of analysis is that the ‘beam’ technologies they use may analyse a smaller mass of material than the minimum recommended test-portion mass that is specified by the RM producer. Theoretical modelling of the mass analysed by a PXRF with a beam size of 8 mm suggested that out of 17 elements where concentrations are quoted for the SdAR series of RMs, only one of these elements could be considered to have a test portion mass exceeding the minimum recommendation of 0.2 g [12]. These estimates of test portion mass are affected by the differential attenuation of fluoresced X-rays by different elements, and for this reason the test portion mass of each element has to be estimated separately.

An experiment was performed to estimate the component of measurement uncertainty caused by heterogeneity in the three SdAR RMs [12]. The objectives of the experiment were to estimate the component of uncertainty caused by heterogeneity for a range of elements and also to compare these estimates between two different PXRF beam sizes (8 mm and 3 mm). The latter objective can be considered as a comparison between uncertainty evaluations, and is discussed below, first using the *F*-test (as used in the original publication), and secondly, using the CI comparison.

The heterogeneity estimated in this experiment may be thought of as having two components: intrinsic variability between the 12 targets, and the additional variability due to sampling very small areas of these targets. Given the experimental design, these two components cannot be separated, so what we compare below is their combined effect at two different beam sizes. Because of the common component, treating the two variance estimates as statistically independent is not strictly valid. The resulting test will be conservative, that is to say there will be a loss of power rather than a tendency to produce spurious significant results, in the same way that doing a two-sample t-test instead of a paired t-test results in a loss of power when there is positive correlation in the paired observations.

***F*-test (assuming normally distributed data)**

Measurements were made in a laboratory with a model Niton XL3t Ultra PXRF mounted in a desktop stand, in order to minimise the effects of additional sampling uncertainties that may occur in the field, caused for example by different placement of the instrument with respect to the sampling target, or variations in environmental conditions. The sampling target in each case was one face of one of 6 pellets made from the SdAR-H1 material. Two measurements were made (on each face) at both beam diameters (8 mm and 3 mm), without moving the pellet, in order to evaluate analytical repeatability. Standard (classical) ANOVA was used to separate the uncertainty due to heterogeneity from the uncertainty due to analytical repeatability. This was achieved by assigning the result of each duplicated measurement of one pellet surface to the sample tier in a nested balanced design (e.g. Sample 1 and Sample 2 in Fig. 1). Each of these measurement values was then replicated so that Analysis 1 = Analysis 2 (Fig. 1), so that the RANOVA program could be used with just 2 levels of variability, instead of the 3 levels it had been designed for. An example of the resulting raw data for Pb is shown in Table 1. Data in this format was input into a previous version of RANOVA (RANOVA2) for each element, and analysed using classical ANOVA. In this way the reported variance of the sample tier corresponds to analytical repeatability, and the variance at the top (Sampling Target) tier is an estimate of the heterogeneity between pellet surfaces, measured across 12 sampling targets i.e. both sides of 6 pellets.

A value U_{het} was defined as the relative standard deviation of the heterogeneity component from the ANOVA, with a coverage factor of 2 (Equation 1):

$$U_{\text{het}} = \frac{2 \times s_{\text{het}}}{x} \qquad \text{Equation 1}$$

Where s_{het} is the standard deviation of the heterogeneity component from ANOVA, and x is the mean of all measurements for the particular element in that RM.

A ratio (U_{hr}) was then defined to quantify the change in U_{het} when the PXRF beam size was reduced from 8 mm to 3 mm (Equation 2):

$$U_{\mathrm{hr}} = \frac{U_{\mathrm{het3mm}}}{U_{\mathrm{het8mm}}} \qquad \text{Equation 2}$$

The F statistic ($U_{\mathrm{het3mm}} > U_{\mathrm{het8mm}}$) was calculated as follows:

$$F = \frac{U^2_{\mathrm{het3mm}}}{U^2_{\mathrm{het8mm}}} \qquad \text{Equation 3}$$

This statistic does not have an F-distribution, because each of the heterogeneity components is estimated using the difference of two mean squares in the corresponding ANOVA. One possible way forward is to approximate its distribution by an F-distribution with degrees of freedom taken from the rows of the ANOVA tables that correspond to heterogeneity.

There were 12 sampling targets for each beam diameter (both surfaces of 6 pellets), corresponding to 11 degrees of freedom (df) for both numerator and denominator. For the single sided hypothesis that the variance of the 3 mm measurement is *greater* than that for the 8 mm, the critical value of $F_{0.05,11,11}$ (i.e. 95 % confidence, for df = 11) is 2.818.

Combining equations 1, 2 and 3 and cancelling the mean of all measurements, we can say that U_{het3mm} is significantly greater than U_{het8mm} if $U_{\mathrm{hr}} > \sqrt{F_{\mathrm{critical}}} = \sqrt{2.818} = 1.67$.

The comparison of U_{hr} with this critical value is shown in Fig. 2, for concentrations of 14 elements measured by PXRF (7 elements expressed as oxides) in the reference material SdAR-H1. The results shown in Fig. 2 suggest that the uncertainty caused by heterogeneity for 6 elements (Ti, Mn, Fe, Nb, Mo, Pb) is sufficient to produce significantly different evaluations of uncertainty when measurements are made at the two different beam diameters. A more cautious analysis that allows for multiple hypotheses testing (14 tests) would apply the Bonferroni adjustment [15]. In that case the alpha-level would be adjusted to 0.05/14 = 0.00357, giving a critical $F_{\mathrm{critical}} = 5.774$ and $\sqrt{F_{\mathrm{critical}}} = 2.40$. The uncertainties caused by heterogeneity for 5 elements (Ti, Fe, Nb, Mo, Pb) would then be found to be significantly different between beam sizes.

Table 1 Example raw data for Pb in SdAR-H1, prepared for analysis by RANOVA2, using the experimental design from Fig. 1. Units are mg kg^{-1}

Pellet number	Pellet surface	S1A1	S1A2	S2A1	S2A2
1	1	3819.3	3819.3	3866.7	3866.7
	2	3893.7	3893.7	3859.7	3859.7
2	1	3850.3	3850.3	3844.3	3844.3
	2	3856.0	3856.0	3859.8	3859.8
3	1	3869.5	3869.5	3884.6	3884.6
	2	3900.0	3900.0	3853.0	3853.0
4	1	3880.6	3880.6	3863.0	3863.0
	2	3851.5	3851.5	3897.8	3897.8
5	1	3836.9	3836.9	3849.1	3849.1
	2	3910.4	3910.4	3890.3	3890.3
6	1	3878.5	3878.5	3853.1	3853.1
	2	3849.6	3849.6	3880.5	3880.5

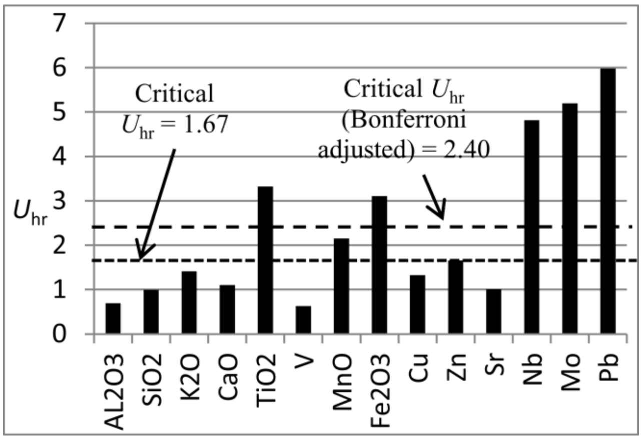

Fig. 2 U_{hr} compared with critical ratios $\left(\sqrt{F_{\mathrm{critical}}}\right)$ for 14 elements in SdAR-H1 [12]

As noted above, these results are based on an approximate *F*-distribution for the test statistic. The approximation will be a good one when the heterogeneity variance component is much larger than the analytical variance component in both the 8 mm and 3 mm cases. Some limited simulations suggest that in this particular example "much larger" can be interpreted as a ratio of 4 or more. Unfortunately this is only true for three elements: SiO2, CaO and Zn. These three apart, the results of the tests should be treated with some caution, and the alternative approach described below is to be preferred.

CI comparison (assuming normally distributed data)

The calculations of CIs for the 8 mm and 3 mm beam sizes were performed using the computer program RANOVA3. Variance estimates and CIs were taken from the *classical* ANOVA. In this case the CIs were calculated using well-established formulae that give approximate CIs for variance components in the case of normally distributed data [9]. The comparison of these CIs is illustrated in Fig. 3.

Using this approach, the heterogeneity of 4 elements (Ti, Nb, Mo, Pb) were shown to produce different evaluations of uncertainty at the two beam diameters. These were also found to be significantly different using the *F*-test, however the elements Mn and Fe were not found to be different in the CI comparison.

Case Study 2: Nitrate in glasshouse grown lettuce

Data for this example is drawn from the Example A1 in the 2nd edition of the Eurachem guide to uncertainty from sampling [5]. An investigation of nitrate concentration in greenhouse grown lettuces was conducted using a standardized sampling protocol, consisting of walking a W-shaped route across each bay of lettuces (a bay contains up to 20,000 heads) and collecting 10 lettuce heads. In this case the sampling target is one bay. Sampling uncertainty was evaluated by taking sample duplicates. The duplicate samples were acquired using a similar protocol with re-orientation of the W-shaped route, and applied to 8 bays (Fig. 4). Each 10-head sample was transported to a laboratory, processed to form a composite sample, and analysed in duplicate by high performance liquid chromatography (HPLC). Measurement results are shown in Table 2, and the results of applying classical and robust ANOVA to these measurements are shown in Table 3. The combined measurement uncertainty (‘Measure’ in Table 3) is obtained by adding the variances of sampling and analysis (Equation 4).

$$u_{\mathrm{Measure}} = \sqrt{u^2_{\mathrm{Sampling}} + u^2_{\mathrm{Analysis}}} \qquad \text{Equation 4}$$

The difference between the classical standard deviation estimate for u_{Sampling} (518 mg kg^{-1}) and the robust estimate (319 mg kg^{-1}) suggests the presence of outlying values in the data, and visual inspection of Table 2 also suggests an outlying sampling difference at sampling target C.

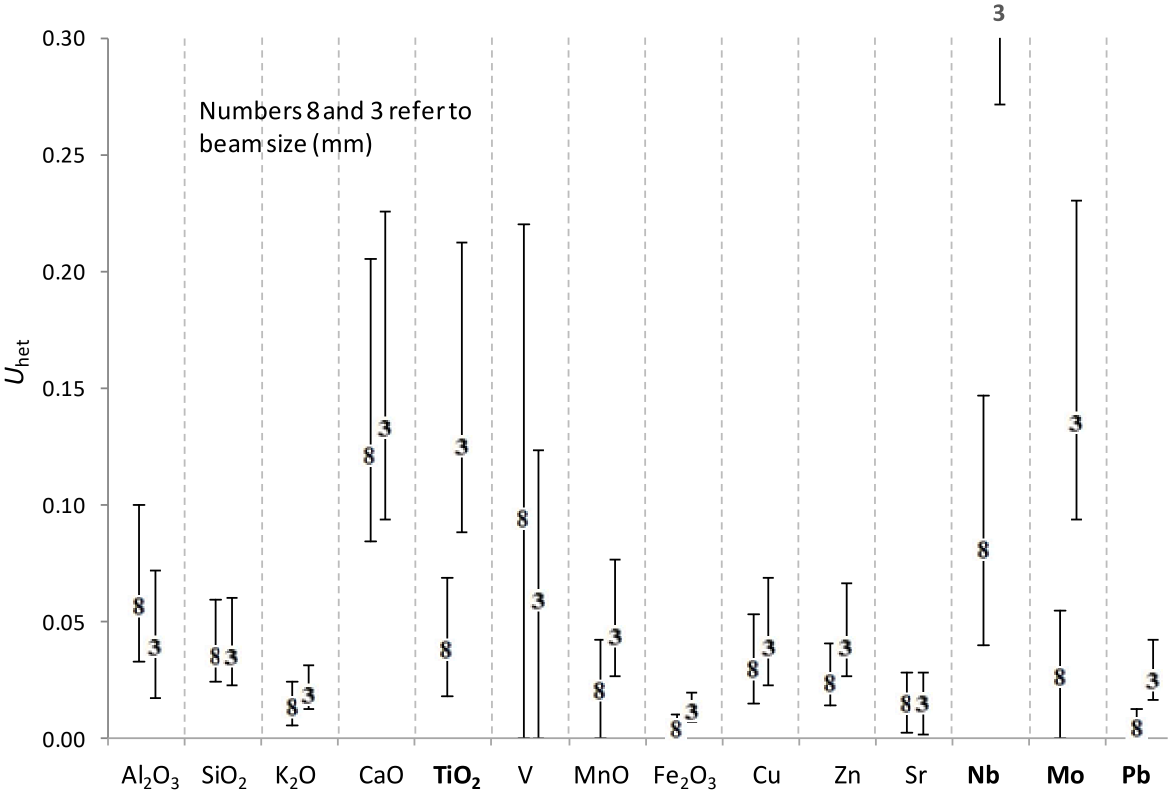

Fig. 3 Comparison of uncertainties using the confidence interval approach, applied to 14 elements (some expressed as oxides) in the SdAR reference material SdAR-H1 as measured using PXRF. Elements with non-overlapping CIs (TiO_2, Nb, Mo, Pb) are shown in **bold**

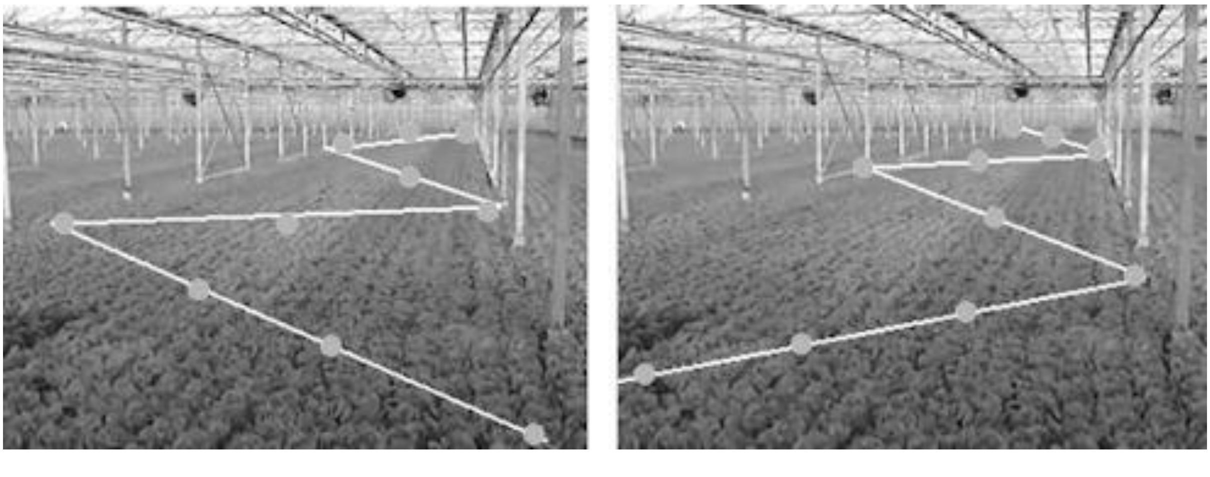

Fig. 4 Sampling protocol (left) specifies taking 10 heads to make a single composite sample from each bay. Duplicate samples (right) were acquired by re-interpretation of the protocol

A CI comparison (introduced in Case Study 1) does not demonstrate a significant difference between the sampling and the combined measurement uncertainties in either the classical or the robust cases. This is because there is overlap of the CIs in both instances. For example the robust value for U'_Sampling is 14.5 (11.3, 32.6), and that for U'_Measure is 16.4 (13.5, 33.4). There is substantial overlap between these two CIs. However there is no overlap between the CIs for U'_Sampling =14.5 (11.3, 32.6) and U'_Analysis = 7.6 (6.2, 9.5), indicating that these estimates of uncertainty are significantly different. The conclusion may be drawn that the uncertainty caused by sampling is the dominant component of the combined uncertainty U'_Measure. This is useful information because it implies that any requirement to reduce the combined uncertainty could most practically be achieved by allocating additional resources to the sampling operation, in preference to the analytical method.

Table 2 Measurements of the concentration (mg kg^{-1}) of nitrate in eight duplicated composite samples of lettuce. Duplicate samples are labelled S1 and S2. Duplicate analyses are labelled A1 and A2

Sampling target	**S1A1**	**S1A2**	**S2A1**	**S2A2**
A	3898	4139	4466	4693
B	3910	3993	4201	4126
C	5708	5903	4061	3782
D	5028	4754	5450	5416
E	4640	4401	4248	4191
F	5182	5023	4662	4839
G	3028	3224	3023	2901
H	3966	4283	4131	3788

Table 3 Summary of results from classical and robust ANOVA performed on 8 duplicated samples of 10 lettuce heads, analysed using RANOVA3 [8]. Standard deviation estimates (*u*) are in units mg kg^{-1}. Lower and upper confidence limits are shown in brackets. *U'* is the expanded (95 %) uncertainty relative to the mean

Classical ANOVA

	Sampling		Analysis		Measure	
SD (or *u*)	518	(334, 1008)	148	(110, 226)	539	(372, 1018)
% of total variance	44.8		3.7		48.4	
U' (Exp Rel 95%)	23.8	(15.4, 46.4)	6.8	(5.1, 10.4)	24.8	(17.1, 46.9)

Robust ANOVA

	Sampling		Analysis		Measure	
SD (or *u*)	319	(248, 720)	168	(137, 209)	361	(298, 736)
% of total variance	22.6		6.3		28.9	
U' (Exp Rel 95%)	14.5	(11.3, 32.6)	7.6	(6.2, 9.5)	16.4	(13.5, 33.4)

For the reasons given in Case Study 1, it would not be recommended to use an *F*-test on variance ratios in this example. However it is possible to perform a limited number of tests directly on the mean-square values that are part of the ANOVA calculation, because these are both statistically independent and their distributions are multiples of chi-squared. Given a balanced experimental design of the form illustrated in Fig. 1, where *I* = the number of targets, *J* = the number of samples taken per target, and *K* = the number of analyses performed per sample (e.g. in Table 2, *I* = 8, *J* = 2, *K* = 2) it is possible to re-calculate the mean-square values from the variances or standard deviations, such as those calculated by RANOVA3.

The mean-square value at the analysis level ($\mathrm{MS_e}$) is simply the analytical variance s_e^2, where s_e is the standard deviation of analysis. So for the **robust** ANOVA results in Table 3:

$\mathrm{MS_e} = s_\mathrm{e}^2 = (168)^2 = 28224$

The mean-square value at the Sample level ($\mathrm{MS_B}$) is calculated as follows, where s_B is the standard deviation of the sampling level:

$\mathrm{MS_B} = Ks_\mathrm{B}^2 + s_\mathrm{e}^2 = 2\,(319)^2 + (168)^2 = 231746$

(Units of standard deviation are mg kg^{-1}, and are squared for units of variance)

The F-ratio $\mathrm{MS_B} / \mathrm{MS_e}$ can be used to test whether the population variance at the sample level (σ_B^2) is significantly different from zero. In this case:

$\mathrm{MS_B}/ \mathrm{MS_e} = 231746/28224 = 8.21$

Degrees of freedom for the respective levels are calculated as $I\,(J-1)$ (numerator) and $IJ\,(K-1)$ (denominator). The F-ratio 8.21 is greater than the critical value $F_{\mathrm{Crit}(0.05,8,16)} = 2.6$, so rejecting the null hypothesis that $\sigma_\mathrm{B}^2 = 0$ at a probability level $\alpha = 0.05$. We can also use the fact that in the particular case of equal variances, i.e. if $\sigma_\mathrm{B}^2 = \sigma_\mathrm{e}^2$ (the population variance at the analysis level), the distribution of:

$$\frac{\mathrm{MS_B}}{\mathrm{MS_e}}\left(\frac{\sigma_\mathrm{e}^2}{\sigma_\mathrm{e}^2 + K\sigma_\mathrm{B}^2}\right) \qquad \text{Equation 5}$$

can be written

$$\frac{\mathrm{MS_B}}{\mathrm{MS_e}}\left(\frac{1}{1+K}\right) = \frac{\mathrm{MS_B}}{\mathrm{MS_e}}\left(\frac{1}{3}\right) \qquad \text{Equation 6}$$

The F-ratio $\mathrm{MS_B}/ \mathrm{MS_e} = 8.21$ is greater than 3 times the critical value (3 × 2.6) = 7.8, so the null hypothesis that $\sigma_\mathrm{B}^2 = \sigma_\mathrm{e}^2$ can also be rejected, and we can conclude that the sampling and analytical variances are significantly different at a probability level $\alpha = 0.05$. This supports the previous conclusion from the comparison of CIs, that the uncertainty caused by sampling is the dominant component of the combined uncertainty.

Case Study 3: Sampling proficiency test (comparison using bootstrapped CIs)

As previously stated, it is generally the case that in chemical measurements, the value of the measurand is effectively the true value of the analyte concentration in the sampling target. For this reason it is now widely agreed that a chemical measurement process begins when the primary sample is taken. Sampling proficiency tests (SPTs) have been performed with the objectives of assessing the performance of different samplers, and using the results of these tests to make improved evaluations of measurement uncertainty [16].

An example of an SPT is given in [16], where 9 participating samplers extracted 6-sample composite samples from blocks that constituted a 20 ton batch of fresh butter produced in a factory, using a hand-held coring device. Each sampler repeated the process in order to produce 2 composite samples that were then analysed gravimetrically to determine the moisture content, expressed as mass fraction (m/m), further details are given in [16]. This is an important parameter for this particular industry, as it is subject to international legislation. Two composite samples were acquired by each sampler in order to evaluate the 'within-sampler' sampling uncertainty. Each of these composite samples was also analysed twice in order to also evaluate the analytical uncertainty, as analytical repeatability. This experimental design therefore conforms to a fully balanced experimental design where individual samplers effectively use the 'duplicate method' (discussed above), but the additional between-sampler variability includes a component of the potential sampling bias from each sampler (Fig. 5). The results of these analyses are shown in Table 4.

The measurements (Table 4) were analysed using robust ANOVA, which is designed to accommodate up to 10 % outlying values, which were evident for samplers I and G [16]. Results of this analysis are shown in Table 5.

The SD, or standard uncertainty (u) for Measurement in Table 5 is the square root of the sum of the resolved variances of Sampling and Analysis, so that:

$$s_{\text{Measurement}}^2 = s_{\text{Sampling}}^2 + s_{\text{Analysis}}^2 \qquad \text{Equation 7}$$

This can be considered as the measurement uncertainty within-sampler, encompassing uncertainty due to sampling and analytical repeatability, averaged across the 9 samplers.

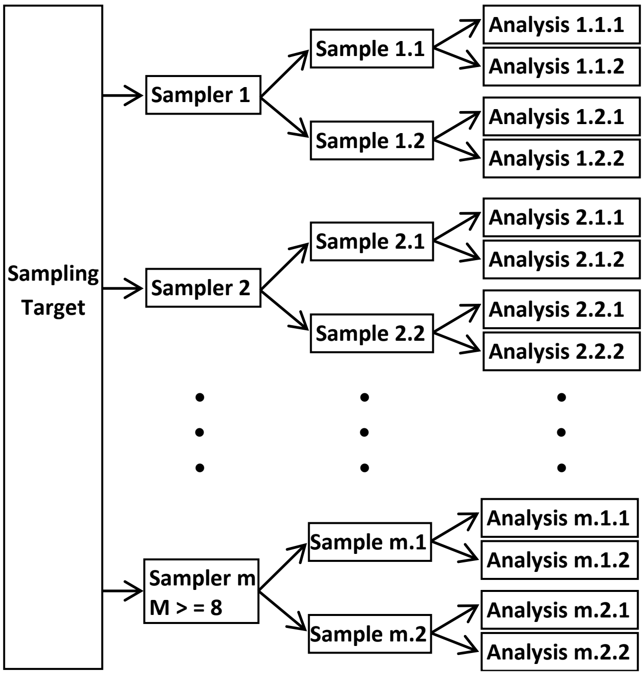

Fig. 5 Balanced design for SPT on a batch of butter (modified from [16]). Note that in this case the different samplers 'Sampler 1'... 'Sampler m' are considered as different sampling targets by the RANOVA program, hence the wording in Table 5

Table 4 Measured concentrations of moisture in butter for the sampling proficiency test, expressed as mass fraction (%)

Sampler (m)	Analysis result m.1.1	Analysis result m.1.2	Analysis result m.2.1	Analysis result m.2.2
A	15.4741	15.4155	15.4972	15.4796
B	15.3655	15.3257	15.3653	15.3373
C	15.4417	15.4069	15.4552	15.4518
D	15.4161	15.4134	15.4486	15.4143
E	15.4085	15.3675	15.4392	15.4060
F	15.4148	15.3876	15.4176	15.3473
G	15.4959	15.4757	15.4853	15.5185
H	15.3673	15.3732	15.3720	15.3427
I	15.3214	15.2779	15.3424	15.3721

Table 5 Results of robust ANOVA on measurements of moisture in butter (expressed as % m/m), extracted from the program RANOVA3. Confidence intervals are enclosed in brackets

Mean	15.4
Total SD (std dev)	0.067 (0.044, 0.089)

	Btn Target[1]	Sampling	Analysis	Measurement
SD (or *u*)	0.060 (0.034, 0.088)	0.013 (0, 0.032)	0.027 (0.023, 0.037)	0.030 (0.024, 0.041)
% of total variance	0.14	3.73	16.13	19.86
U' (Exp Rel 95%)		0.17 (0, 0.41)	0.35 (0.30, 0.47)	0.39 (0.31, 0.53)

[1] Between Target

The Total SD in Table 5 is the square root of the total variance, which is the sum of the variances for between-sampler ('Btn Target' in Table 5), within-sampler ('Sampling' in Table 5), and Analysis, so that:

$$s^2_{\text{Total}} = s^2_{\text{Btn Sampler}} + s^2_{\text{Measurement}} \qquad \text{Equation 8}$$

This includes the between-sampler component of uncertainty due to variance caused by different amounts of potential sampling bias between different samplers. It can be converted into an expanded relative uncertainty as follows:

$$U' = 200\left(\frac{s_{\text{Total}}}{\bar{x}}\right) \qquad \text{Equation 9}$$

Where $\bar{x}$ is the mean of all measurements. The results, including CIs, are summarised in Table 6.

Table 6 Summary of uncertainty on SPT data with bootstrapped confidence intervals

	s	U'
Measurement (including between and within-sampler)	0.067 (0.044, 0.089)	0.87 (0.57, 1.16)
Measurement (including within-sampler only)	0.030 (0.024, 0.041)	0.39 (0.31, 0.53)
Analytical only	0.027 (0.023, 0.037)	0.35 (0.30, 0.48)

Table 6 shows that the measurement uncertainty (including between-sampler and within-sampler, $U' = 0.87$) exceeds the measurement uncertainty (including within-sampler only, $U' = 0.39$) by a factor of 2.2. The study suggests that the larger value is likely to be a more realistic estimate of the uncertainty in this case because it includes the usually overlooked contribution from potential bias between different samplers.

Using the recently published approach of estimating confidence intervals on the results of robust ANOVA by bootstrapping, it is now possible to test the significance of this interpretation. It can be seen that the lower CL on U' for multi-sampler (0.57) exceeds the upper CL on U' for within-sampler (0.53) (Table 6). Therefore, there is no overlap between the 95 % confidence intervals on these uncertainties, so we can conclude that they are significantly different, and that between-sampler bias is a factor contributing to the overall uncertainty. The same conclusion can be obtained by applying an F-test to the mean-squares, using a similar approach to that described in Case Study 2. This conclusion is an approximation only, because the variances used to construct the mean squares in what follows were estimated using a robust algorithm, instead of classical ANOVA. If MS_A = the mean square at the between-sampler level, MS_B = the mean square at the within-sampler level, the F-ratio MS_A / MS_B can be used to test whether the population variance at the between-sampler level (σ^2_A) is significantly different from zero. Using the standard deviation values from the robust ANOVA results shown in Table 5, where J = the number of samples taken by each sampler, K = the number of analyses performed per sample, s_A = the standard deviation at the between-sampler level, s_B = the standard deviation at the within-sampler level, and s_e = the standard deviation at the analytical level, it is possible to calculate the mean-squares from these standard deviations:

$$\text{MS}_\text{B} = Ks^2_\text{B} + s^2_\text{e} = 2\,(0.013)^2 + (0.027)^2 = 0.00110$$

$$\text{MS}_\text{A} = JKs^2_\text{A} + \text{MS}_\text{B} = 4\,(0.060)^2 + 0.00110 = 0.0155$$

$$\text{MS}_\text{A} / \text{MS}_\text{B} = 14.1$$

Degrees of freedom for the respective levels are calculated as $I - 1$ (numerator) and $I\,(J - 1)$ (denominator). The F-ratio 14.1 is greater than the critical value $F_{\text{Crit}(0.05,8,8)} = 3.73$, so rejecting the null hypothesis that $\sigma^2_\text{A} = 0$, confirming that the between-sampler uncertainty (corresponding to 'Btn Target' in Table 5) is a significant factor in the overall uncertainty. Confirmation of this is useful information because it indicates that uncertainty estimated by the duplicate method alone (i.e. the usual method using just a single sampler) is likely to be an underestimate. This is because it does not include any component of sampling bias. In practice the greater costs of using a multi-sampler approach may only be justifiable where the consequential cost of misclassifying the target is also high [17].

Discussion

Two approaches for comparing two uncertainties have been demonstrated in the three preceding case studies. Case study1compared an approximate *F*-test with the CI comparison. In this example the CIs were calculated for classical ANOVA by RANOVA3 using formulae that apply to normally distributed data. It was found using the *F*-test that 6 out of 14 elements were sufficiently heterogeneous in the sampling target to cause a significant difference in the uncertainties when using a PXRF instrument set at two different beam sizes (Fig. 2). When the CI comparison procedure was used the difference was significant for 4 of these elements, but the uncertainties for Fe and Mn were not found to be significantly different (Fig. 3).

Both of these approaches are approximate. The safe choice is the comparison of confidence intervals, because the approximation used to construct the intervals is generally a good one. The approximate *F*-test can only be relied on in some circumstances, essentially when the variance component of interest is much larger than the ones that are below it in the ANOVA hierarchy and whose contribution needs to be subtracted in its estimation. When the *F*-test is valid, it has more power than the comparison of confidence intervals.

Case study 2 provides an example of comparing uncertainty components in order to establish whether the sampling or analytical components are dominant in the overall (combined) uncertainty. It is not valid to use an *F*-test directly on the variances calculated by nested ANOVA because they are not statistically independent. It is possible to perform limited tests on the mean-square values that are part of the ANOVA calculations. Where necessary these can be re-calculated from the variances or standard deviations, and the parameters of the experimental design.

Case study 3 is an example of the comparison of uncertainties when the robust analysis of variance is used, which is the recommended approach if measurement data is suspected to contain a small proportion (< 10 %) of outlying values [5]. The program RANOVA3 provides CIs for robust ANOVA by using a bootstrapping approach. Non-overlap of the CIs on two uncertainties implies a significant difference.

The case studies and methods described in this paper apply to Type A uncertainty evaluations, that is uncertainties that have been evaluated by statistical analyses of measured quantity values [4]. These methods have not been applied or tested on uncertainties that have been wholly or partially quantified using Type B evaluations.

Conclusion

In some situations it is useful to compare evaluations of measurement uncertainty, for example to compare the fitness for purpose of two different analytical methods in chemistry. It has also been shown possible to compare the different contributions of uncertainty that combine to make the overall measurement uncertainty.

In situations where the *F*-test of a variance ratio is valid, it is the preferred option, but there are many common situations where it is not valid. If the only issue is non-normality, then an *F*-test applied to the ratio of robustly estimated variances is the best option. If the lack of validity arises from the fact that the variances have been estimated by the subtraction of mean squares in one or more ANOVAs, an approximate *F*-test is still an option if the variance component of interest is several times larger than the one(s) below it in the ANOVA hierarchy. Failing any of these, it is always possible to compute two separate confidence intervals, either using standard formulae in the normal case or bootstrapping in the non-normal case, and check for overlap. This test has relatively low power, but this may not matter if the variances are very different. A further possibility would be to compute bootstrap confidence intervals for the variance ratio in the difficult cases, but this would require new software to be developed.

Robust ANOVA and CI calculations are available in the computer program RANOVA3, available from the website of the Royal Society of Chemistry [8].

References

1. Ellison SLR, Williams A (Eds.) Eurachem/CITAC guide: Quantifying uncertainty in analytical measurement, Third edition, (2012) ISBN 978-0-948926-30-3 https://www.eurachem.org/index.php/publications/guides/quam
2. International Organization for Standardization. ISO/IEC Guide 98-4:2012(en) Uncertainty of measurement — Part 4: Role of measurement uncertainty in conformity assessment. ISO: Geneva, Switzerland https://bbn.isolutions.iso.org/obp/ui#iso:std:iso-iec:guide:98:-4:ed-1:v1:en
3. International Organization for Standardization. ISO/IEC Guide 98-3:2008(en) Uncertainty of measurement - Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) . ISO: Geneva, Switzerland https://bbn.isolutions.iso.org/obp/ui#iso:std:iso-iec:guide:98:-3:ed-1:v2:en5
4. International Organization for Standardization. ISO/IEC Guide 99:2007(en) International vocabulary of metrology — Basic and general concepts and associated terms (VIM). ISO: Geneva, Switzerland https://bbn.isolutions.iso.org/obp/ui#iso:std:iso-iec:guide:99:ed-1:v2:en
5. Ramsey MH, Ellison SLR, Rostron PD (eds.) (2019) Eurachem/EUROLAB/ CITAC/Nordtest/AMC Guide: Measurement uncertainty arising from sampling: a guide to methods and approaches. Second Edition, Eurachem (2019). ISBN (978-0-948926-35-8) https://www.eurachem.org/index.php/publications/guides/musamp
6. Lyn JA, Ramsey MH, Coad S, Damant AP, Wood R, Boon KA (2007) The duplicate method of uncertainty estimation: are eight targets enough? Analyst 132, 1147-1152 https://doi.org/10.1039/B702691A
7. AMC (1989) Robust statistics - how not to reject outliers. Part1, basic concepts. Analyst 114:1693-1697 https://doi.org/10.1039/AN9891401693
8. AMC (2020) RANOVA3 computer program https://www.rsc.org/membership-and-community/connect-with-others/through-interests/divisions/analytical/amc/software/
9. Rostron PD, Fearn T, Ramsey MH (2020) Confidence intervals for robust estimates of measurement uncertainty. Accreditation and Quality Assurance 25:107–119 https://doi.org/10.1007/s00769-019-01417-4
10. Schenker N, Gentleman JF (2001) On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. The American Statistician 55 (3) 182-186 https://doi.org/10.1198/000313001317097960
11. Horwitz W (1990) Nomenclature for sampling in analytical chemistry (Recommendations 1990) International Union for Pure and Applied Chemistry, Pure and Applied Chemistry 62, 1193–1208 https://doi.org/10.1515/iupac.62.0018
12. Rostron PD, Ramsey MH (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. Geostandards and Geoanalytical Research 41(3) 459-473 https://doi.org/10.1111/ggr.12162
13. Webb PC, Thompson M, Potts PJ, Wilson SA (2014) GeoPT35A — An international proficiency test for analytical geochemistry laboratories — Report on supplementary round 35A (metalliferous sediment, SdAR-H1) https://www.google.com/search?client=firefox-b-d&q=GeoPT35AReport.pdf
14. Ramsey MH, Wiedenbeck M (2017) Quantifying isotopic heterogeneity of candidate reference materials at the picogram sampling scale. Geostandards and Geoanalytical Research 42 (1) 5-24. ISSN 1751-908X https://doi.org/10.1111/ggr.12198
15. Jean OJ (1961) Multiple comparisons among means. Journal of the American Statistical Association 56 (293) 52-64 https://doi.org/10.2307/2282330
16. Ramsey MH, Geelhoed B, Wood R, Damant AP (2011) Improved evaluation of measurement uncertainty from sampling by inclusion of between-sampler bias using sampling proficiency testing. Analyst 136, 1313-1321

https://doi.org/10.1039/C0AN00705F

17. Thompson M, Fearn T (1996) What exactly is fitness for purpose in analytical measurement? Analyst, 121, 275–278 https://doi.org/10.1039/AN9962100275