Results from an Amino Acid Racemization Inter-Laboratory Proficiency Study, Part 2; Measurement Uncertainty Evaluation

J Powell, K Penkman, J Cussens, N MacLeod, M Collins

ABSTRACT

In the first part of this paper we looked at the evaluation of proficiency test (PT) data, both in terms of the precision (random error effects) of submitted results but specifically in terms of the relative bias (systematic error effects) when compared against the assigned value for a given analyte, determined as the consensus of submitted results. Due to absence of defined reference materials, it is not currently possible to assess or correct for bias in routine amino acid racemization (AAR) analysis. This does not affect the technique's application as a relative dating method but hinders comparability of data between laboratories and limits its wider applications such as the development of inter-regional or even global aminostratigraphies and palaeoclimte reconstruction. Because of these difficulties, measurement uncertainty (MU) in AAR geochronology is currently reported using precision estimates. Precision and bias are both essential elements of "Top-down" approaches to uncertainty determination. However, without evidence of the absence of bias, uncertainty estimates based solely on precision are probably currently being underestimated.

In Part 2 we now consider how the uncertainty of a measurement result is the product of a hierarchy of random and systematic error effects, referred to as "The ladder of errors", and how different approaches to MU evaluation are able to account for different levels of error. Proficiency test data is recognised as providing a valuable contribution to uncertainty measurement yet there is little information available and few examples on how this should actually be done. Here we evaluate uncertainty estimates using the results of the proficiency test study, using four different approaches, i) solely as precision estimates from replicate data, ii) using precision and bias data to give a combined estimate for a single PT result, iii) using precision and bias data to give a combined estimate over a series of PT results, iv) using ANOVA to derive reproducibility precision as an estimate of the overall uncertainty.

1 INTRODUCTION

1.1 Measurement Uncertainty (MU)

For the majority of users of analytical data, the information provided by a measurement value is assumed to be the real value or true value. However, a single measurement or even a group of measurements simply represents one (or several) of many possible values for the given measurand. If the same analysis was to be repeated again by the same person or even a different person then the results may be very slightly, or sometimes even quite significantly different. The result is thus only a representation of our best estimate given the limitations of the equipment, conditions, expertise etc, the true value remains unknown. For this reason it is necessary to assess the dispersion of other possible values of our estimate for the same measurand and report it alongside our measurement result. This parameter is known as the measurement uncertainty and provides a quantitative expression of the level of doubt associated with a reported result.

In 1978, the Comite International des Poids et Mesures (CIPM), recognised the lack of uniformity in the handling of uncertainty measurement. After fifteen years of international collaboration, a set of common fundamental principles were established with the publication of the first authoritative document; the Guide to the Expression of Uncertainty in Measurement in 1993.

The Guide or GUM as it has come to be known is still commonly accepted as the international definitive guidance document for uncertainty measurement, although since then various supporting documents have been written to assist in its interpretation and implementation at bench level and several other alternative methodological approaches have been proposed. However it was the publication of the GUM that has resulted in the global consensus on reporting uncertainty associated with measurements and has enabled comparison and standardisation of those results in calibration, accreditation, and analytical service around the world.

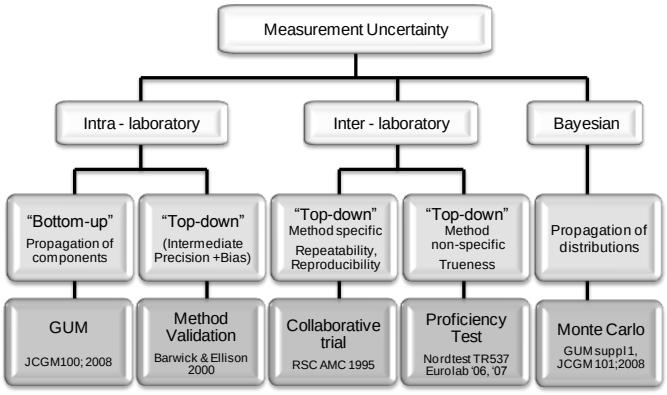
According to the VIM 2.26 (JCGM 200:, 2008), measurement uncertainty (MU) is defined as a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used". MU includes both systematic and random error effects. Contributions are characterized by standard deviations and may arise from statistical distributions of repeated measurements (Type A components) or evaluated from probability density functions based on experience or other information such as certificates and specification sheets (Type B components). It is important for individual components (standard uncertainties) to be evaluated and expressed in a standardized way so that they can be more easily combined using normal procedures for combining variances. The recommended expression of MU is termed the Expanded uncertainty (U), and requires the combined uncertainty estimate (u_c) to be multiplied by a value known as the coverage factor, (k). The coverage factor reflects the level of confidence or probability level required for the measurement result. An expanded uncertainty estimate should accompany every measurement result, and may be expressed either as an absolute value or as a relative percentage, and specify the value of k used (EURACHEM / CITAC, 2000, JCGM 100:, 2008).

1.2 "Bottom-up", the GUM approach to MU.

Initially the recommended method for determining the uncertainty due to error was to carry out an exhaustive audit of individual standard uncertainty contributions from every part of the measurement system. This approach, referred to as the "bottom-up" approach, was often criticized for underestimating the combined standard uncertainty as for all but the simplest analytical models, error effects often went unaccounted for, took too long and was too labour intensive. Accurate determination of MU must encompass all error contributions at all stages of the measurement system. Consequently an alternative method known as the "top-down" approach was introduced. This recognizes that individual random and systematic error influences are reflected in the results of an analytical measurement and looks at evaluating the overall contribution, looking down from above as it were. A summary of the different approaches adopted for uncertainty evaluation is shown in Figure 1.

The next three methods can all be described as "top-down" approaches to uncertainty evaluation; single or intra-laboratory method validation, the collaborative trial or inter-laboratory method validation and proficiency testing. Each will now be considered in turn.

Figure 1; Routes for measurement uncertainty determination



1.3 "Top-down" approaches to MU.

For any analytical measurement there are a number of sources of error, composed of both random and systematic error effects. These can be classified as a hierarchy, sometimes referred to as "the ladder of errors" (Thompson, 2000), see Figure 2 Strictly speaking we can't actually know the true value, just our best estimate with a stated region representing the range of other possible determinations of the measurand's value, under the influence of error.

The Royal Society of Chemistry's Analytical Methods Committee (1995), help to explain that any measurement value (x), is the result of the effect of the error contributions on the true value (μ), thus;

$$x = \mu + \delta_{method} + \delta_{lab} + \delta_{run} + \varepsilon \qquad (1)$$

Where δ_{method} is the method bias, δ_{lab} is the laboratory bias, δ_{run} is the run bias and ε is the random error component or repeatability. Therefore, in order to express a measurement result correctly, our estimate of these error effects must be reflected in the final expression of the result, that is; our best estimate of the true value with an associated uncertainty. Therefore we now have;

$$\mu = x \pm u(x) \qquad \text{and} \qquad (2)$$

$$u(x) = \sqrt{u(\varepsilon)^2 + u(run)^2 + u(lab)^2 + u(method)^2} \qquad (3)$$

Where *u(method)* is the uncertainty due to method bias, *u(lab)* is the uncertainty due to laboratory bias, *u(run)* is the uncertainty due to run bias and *u(ε)* is within-run repeatability uncertainty due to random error effects.

1.3.1 Intra-Laboratory Method Validation approach.

It was recognised that much of this information was already being determined by laboratories carrying out precision and bias experiments as part of routine internal method validation procedures and so required little additional time, effort or expense on the parts of the analysts. Precision assess random error effects, expressed as the within run (repeatability) or between run standard deviation, whilst bias assess systematic error effects.

Occasionally there are situations where internal control is all that is required and absolute accuracy is not necessary (or possible), then method and laboratory bias components can be ignored, such as in factory production monitoring (AMC, 1995). The inability of AAR geochronology laboratories to currently evaluate bias routinely due to the absence of defined reference materials is another scenario where the uncertainty on the measurement results is simply expressed as a precision estimate. In this case it is the combined repeatability uncertainty and the run to run uncertainty, expressed as standard deviations;

$$u(x) = \sqrt{u(\varepsilon)^2 + u(run)^2} \qquad (4)$$

Repeatability uncertainty, $u(\varepsilon)$, includes weighing and measuring errors, heterogeneity of sample portions, preparation and extraction stages and random instrumental effects. The run effect, $u(run)$, reflects day to day differences in the analytical system such as reagents used, the analysts, temperature effects and even instruments. These are seen as systematic error influences on all the samples in the whole run. Therefore for a single run, the run bias is fixed, but when viewed from a higher level, i.e., over a number of successive runs, the run to run biases can be seen as a random variable and characterized by a standard deviation describing the run to run precision, see Figure 3(a). Both of these effects are usually determined from precision experiments carried out as part of a single-laboratory's method validation, and expressed as standard deviations, taking to account matrix and concentration differences (Barwick and Ellison, 2000a). Further, when taken together, they can provide an indication of overall precision or reproducibility standard deviation (s_R) for a single laboratory. Reproducibility is most often associated with inter-laboratory precision estimates, therefore to make the distinction between within or intra-laboratory precison (s_{Rw}) it is sometimes referred to as the Intermediate precision. Note, when a measurement result is expressed as a mean of repeated measurements, the standard uncertainty should be more correctly expressed as the standard error of the mean (also referred to as the standard uncertainty of the mean or standard deviation of the experimental mean) , Thus;

$$s_{Rw} = \sqrt{{s_\varepsilon}^2/n + {s_{run}}^2} \qquad (5)$$

Where, $u(\varepsilon) = {s_\varepsilon}^2/n$ is the standard uncertainty of the mean, (where *(n)* is the number of replicates) and $u(run) = s_{run}$.

To date, AAR uncertainty estimates have been limited to simple expressions of precision such as the within run standard deviation s_ε or possibly the between run standard deviation s_{run}. Whilst individual laboratory precision and consistency may be all that is required for the method's application as a relative dating technique, in the absence of bias control, correction or bias uncertainty contributions, data will not be reproducible between laboratories at this level. Further, any numerical ages thus determined may grossly underestimate uncertainty estimates and associated confidence intervals reported up till now (Westaway, 2009).

Bias experiments performed as part of a single laboratory trueness evaluation will provide important information about an individual laboratory's combined method and laboratory bias. Where suitable matrix-matched certified reference materials are available, the difference between the mean result ($\bar{x}$) of repeated analyses of a CRM and the certified value (X_{ref}) will provide an estimate of the combined bias. When evaluating bias, it is usual to assess its significance using a student's t-test [ref]. Any measurement result should normally be corrected for significant bias, unless otherwise stated for example in the case of

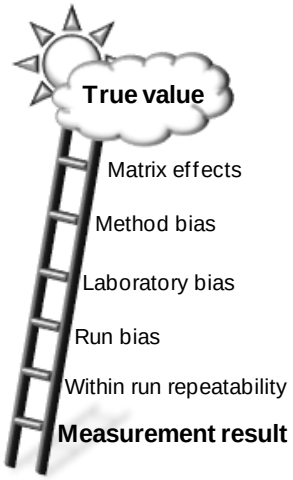


Figure 2; The Ladder of Errors. For each rung of the ladder, the measurement result accumulates increasing uncertainty contributions.

empirical methods. Bias is usually corrected for by the use of reference materials such as an internal standard, spiking and/or calibration.

$$single\ laboratory\ bias_{method+lab} = (\bar{x} - X_{ref}) \quad (6)$$

The standard uncertainty of the bias, *u(bias)*, is then the uncertainty associated with the mean of replicate values, $u(\bar{x})$, i.e.; the standard deviation of the mean; $s_{\bar{x}}/\sqrt{n}$, plus the uncertainty of the reference material or CRM, $u(X_{ref})$. Thus

$$u(bias) = \sqrt{u(\bar{x})^2 + u(X_{ref})^2} \quad or \quad (7a)$$

$$u(bias) = \sqrt{{s_{\bar{x}}}^2/n + u(X_{ref})^2} \quad (7b)$$

Combining equations 5 and 7b, we now have an expression for the combined uncertainty for a single laboratory;

$$u_c = \sqrt{{s_\varepsilon}^2/n + {s_{run}}^2 + {s_{\bar{x}}}^2/n + u(X_{ref})^2} \quad or \quad (8a)$$

$$u_c = \sqrt{{s_{Rw}}^2 + u(bias)^2} \quad (8b)$$

It should be noted that where the uncertainty of the CRM is small, perhaps less than 10% of the analytical uncertainty, it will have a negligible effect on the overall uncertainty of the bias and can be omitted [ref]. It should also be noted that when both precision and bias are being determined in the same analytical run, there is the risk of double counting uncertainty due to random error effects since $u(\varepsilon) = u(\bar{x})$ [ref]. If repeatability precision is small compared to the between run component, then again, this need not be of concern, otherwise the combined uncertainty will need to be adjusted appropriately.

It is therefore conceivable that under certain circumstances, the two bias uncertainty components may be omitted from the combined uncertainty calculation, which would simply then be equivalent to the intermediate precision estimate;

$$u_c = \sqrt{{s_\varepsilon}^2/n + {s_{run}}^2 + \cancel{{s_{\bar{x}}}^2/n} + \cancel{u(X_{ref})^2}} \equiv \sqrt{s_{Rw}^2} \quad (9)$$

Consequently any variability due to individual laboratory/method bias in this context, will be reflected in the precision estimate.

However, it was observed that often a pre-occupation with intra-laboratory precision neglected important contributions from the higher order laboratory and method biases (RSC Analytical Methods Committee, 1995), when determining performance characteristics for a method. Consequently a double edged approach is generally advocated (refs) when determining uncertainty estimates as part of method validation requirements; that is both the intra-laboratory estimates together with a method specific inter-laboratory collaborative trial in order to obtain inter-laboratory precision values.

1.3.2 Inter-Laboratory Method Validation or Collaborative trial approach

A collaborative trial is a method specific inter-laboratory study whose purpose is to characterize the performance of an analytical method for specified materials and often at specified analyte concentrations, across different laboratories. Because it is method prescriptive, and the best estimate of the true value is derived as the consensus of participating results, method bias and its uncertainty contribution are zero. Thus;

$$u(x) = \sqrt{u(\varepsilon)^2 + u(run)^2 + u(lab)^2} \quad (10)$$

Individual laboratory biases are incorporated into the run or between laboratory precision estimates, in the same way as the run bias became a random variable in the previous section. So whilst laboratory bias for an individual laboratory may be a fixed value in any given run, when biases for several laboratories are considered, from a higher level, these too can be viewed as a random variable and once again, described using precision estimates (refs) (see Figure 3(b)),. Precision estimates in a collaborative trial are derived using a one-way analysis of variance, abbreviated to ANOVA.

Full details on the calculations of S_R, S_L and S_r can be found in (ISO 5725, 1994, ISO 21748, 2010). However, in summary, precision estimates are calculated using ANOVA, thus;

$$s_r = \sqrt{MS_w} \quad (11)$$

$$s_L = \sqrt{(MS_b - MS_w)/n} \quad (12)$$

Where; MS_w within groups mean square and MS_b is the between groups mean square and *n* is the number of replicates.

In this context, the standard deviation of repeatability, s_r, is the equivalent of the within laboratory precision estimate, and the between laboratory precision, s_L, is the run plus the laboratory components. Thus;

$$s_r = s_\varepsilon \quad (13)$$

$$s_L = \sqrt{{s_{run}}^2 + {s_{lab}}^2} \quad (14)$$

The overall expression of uncertainty is the reproducibility standard deviation, s_R, and incorporates both the above terms;

$$s_R = \sqrt{{s_r}^2 + {s_L}^2} \quad (15)$$

So, it can be seen that the reproducibility standard deviation is also equivalent to the intermediate precision estimate from single laboratory precision studies, but with the added laboratory effect, i.e.;

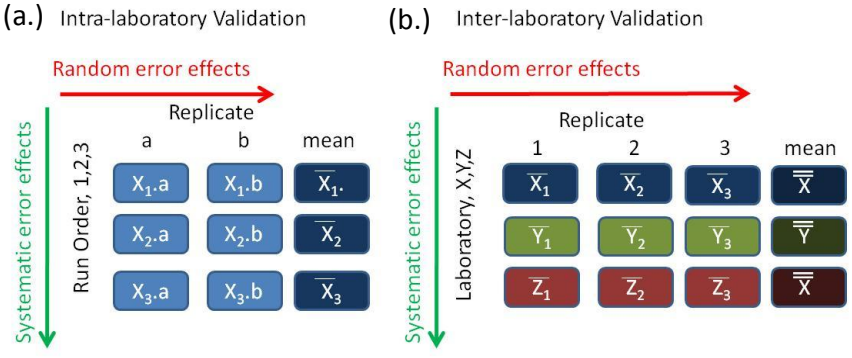


Figure 3; Relationship between Intra and Inter-laboratory Random and Systematic Error Effects.. (a) The systematic uncertainty within a single run (X_1) is a fixed level, but when viewed as one of a number of successive runs (X_1, X_2, X_3 etc), it becomes a random variable with variance $\boldsymbol{s^2_{run}}$. (b) Similarly for a particular laboratory (X), the bias is fixed but when seen as one of a number of laboratories (X, Y, Z etc), again it becomes a random variable with variance $\boldsymbol{s^2_{lab}}$

$$s_R = \sqrt{{s_{Rw}}^2 + {s_{lab}}^2}. \quad (16)$$

In this context, in the absence of method bias, the reproducibility standard deviation (S_R), the overall uncertainty of the analytical system and can be used directly in place of the combined standard uncertainty, providing that the laboratory's own uncertainty estimates can equal or better the published repeatability value. Thus;

$$u_C = \sqrt{{s_{Rw}}^2 + u(bias)^2} \cong s_R \quad (17)$$

Because collaborative trials are regarded as giving unbiased precision estimates, reproducibility values are often used as a benchmark in uncertainty determination since they describe any laboratory's expectation of uncertainty for the stated method when applied to a specified matrix, usually at a given analyte concentration. It is therefore often helpful to compare a laboratory's own uncertainty estimates with those from a collaborative trail. If the laboratory's individual estimates are much smaller than the reproducibility precision estimates, then it is likely that significant sources of uncertainty have been overlooked by the laboratory.

However, collaborative trials on their own generally fail to give the whole picture as they don't include the higher levels of uncertainty at the top of the ladder; the method bias, (unless a CRM was used as the candidate test material), or matrix effects.

1.3.3 Inter-Laboratory Proficiency test approach

Proficiency testing (PT) focuses on the evaluation of analytical trueness or bias as an indicator of accuracy and performance. This is especially valuable in the absence of CRMs, other reference materials or collaborative trail data. Like a collaborative trial, the best estimate of the true value is often taken as the consensus of submitted results, however unlike a collaborative trial, a proficiency test is method non-specific and participants are encouraged to use their routine methods of analysis. Thus a proficiency test can provide valuable information on individual laboratory bias by comparing a participants' submitted result with the consensus or other reference value, but also on method bias if sufficient laboratories submit method specific results, and if evaluated across different matrices, even matrix effects. Proficiency test results therefore potentially and uniquely reflect all of the errors of analysis combined, i.e.; the worst case scenario. However, these additional levels of uncertainty are not generally utilized.

Method bias and it's uncertainty, relate to the level of agreement *between different methods*. Generally speaking, it will be unlikely that different methods will be employed by a single laboratory for the same measurement determination and therefore tends not to be evaluated routinely. (Note, that this is not the same as analytical bias which may be associated with a single method analysis and detected for instance, using bias experiments during method validation and later corrected). The exception to this would be in the case of empirical methods where there may be some interest in knowing the extent to which methods vary between each other.

Generally, a validated method is defined by its matrix or class of materials and perhaps a concentration range, where the conditions of validation hold. However the effect of matrix variation within the defined class of materials is little understood, (eg, terrestrial or marine mollusk shells, egg shell, coral etc, within the class of calcitic biominerals). In principle, this is not difficult to evaluate (refs). However, the difficulty arises in determining the individual biases in the absence of reference values if matrix-matched CRMs are not available. For this reason, validated methods will tend to be matrix and or analyte specific, often with a defined concentration range which together describes the scope of validation and avoids the need to evaluate higher level uncertainties.

The use of proficiency test data in determining measurement uncertainty will be further expended on throughout the rest of this paper.

Thus it can be seen, that the uncertainty from a set of repeated measurements ($u(\bar{x})$), in reality is far from simply being just the standard deviation of the results on their own, or even the standard deviation (uncertainty) of the mean, $s_{\bar{x}}/\sqrt{n}$, but should reflect analytical uncertainty contributions resulting from between run, laboratory, method and even matrix and concentration bias. Often, single-laboratory method validation is insufficient to characterize all of these components, requiring a further inter-laboratory collaborative trial to account for the laboratory bias (expressed as the inter-laboratory precision; see text). Proficiency test data, is usually seen as being limited to evaluations of trueness however, quite uniquely, it can also provide insight into the combined effect from all uncertainty sources.

1.3.4 Monte Carlo approach

One final approach to MU evaluation shown in Figure 1 but has not so far been mentioned and without which this section would not be complete, is that which uses a Bayesian statistical approach to model the propagation of theoretical uncertainty distributions, known as Monte Carlo methods. Although these techniques are not new, they have only been introduced into the realms of measurement uncertainty estimation relatively recently as a supplement to the original GUM document [ref]. However, this method has not been utilized in the context of the current study and interested readers are encouraged to refer to the GUM supplement for further information.

1.4 Expanded Uncertainty (*U*).

The final step in determining the measurement uncertainty and for a measurement to be reported correctly, an expanded uncertainty (*U*) should be given, where the standard uncertainty is multiplied by the relevant coverage factor (*k*), (EURACHEM / CITAC, 2000, JCGM 100:, 2008).thus;

$$U = u(\bar{x}) \times k \quad (18)$$

As a generalization, $k = 2$ is often used by laboratories to represent a 2 standard deviation or approximately 95% confidence interval. For large data sets, perhaps where n=30 or more (Currell and Dowman, 2005), where the distribution of mean values conform with the expectation of normality, this would be acceptable. The 2 standard deviation upper and lower confidence limits either side of the true or population mean, μ, is therefore;

$$\mu - \left[2 \times \frac{\sigma}{\sqrt{n}}\right] \quad to \quad \mu + \left[2 \times \frac{\sigma}{\sqrt{n}}\right] \quad (19)$$

However for decreasing values of *n*, the characteristic bell shaped curve of the normal distribution broadens and flattens reflecting the reduced confidence in the value $\bar{x}$ as the best estimate of the true mean μ, and our uncertainty estimate increases. To compensate for the use of the sample standard deviation, *s*, rather than the population standard deviation σ, k=2 is replaced by the critical *t*-value as a correction term. The value of *t* depends on the value of *n* and the required level of confidence and can be read from any two-tailed *t*-table found in statistical texts. Figure 4 illustrates the relationship between a normal and a t distribution. Thus for n=4 (degrees of freedom=4) at 95% confidence level (α=0.05), t=3.18 compared to the original value of $k=2$. For a pair of replicates; n=2, df=1, t=12.7 and the expanded uncertainty becomes over six times larger than otherwise predicted if k=2! Thus the range in which the true value lies with 95% confidence broadens and becomes;

$$\bar{x} - \left[t_{(2,o.o5,df)} \times \frac{s}{\sqrt{n}}\right] \quad to \quad \bar{x} + \left[t_{(2,o.o5,df)} \times \frac{s}{\sqrt{n}}\right] \quad (20)$$

Results should be expressed as; $\bar{x} \pm U$ (at 95% confidence, using $k=2$, or $k=t_{(2,0.05,df)}$ and *n=the number of data points*).

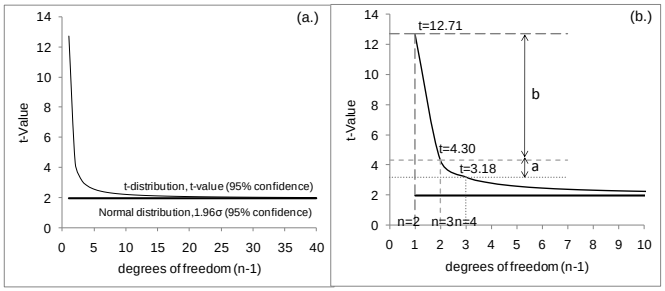


Figure 5; Critical t-values compared to a Normal distribution at 95% confidence level. (a) Shows correction values required to return a t-distribution back to Normal and it's dependence on sample size (degrees of freedom). (b) illustrates how expanded uncertainty can be significantly improved (distance 'b' giving the greatest gain compared to distance 'a') by increasing the sample size from n=2 to n=3.

Because a combined uncertainty brings together uncertainty contributions from different sources, determining *k* becomes a little more tricky when there is no single value for the degrees of freedom. One approach is to calculate an effective degree of freedom using the Welch-Satterthwaite formula where the effective degree of freedom is less than or equal to the sum of the individual values, i.e.; ($v_{eff} \leq \sum v_i$) . The use of this equation is covered in detail in Annex G of the Guide to Uncertainty Measurement or "GUM"; (JCGM 100:, 2008).

However, Eurachem make the following recommendation; "*Where the combined standard uncertainty is dominated by a single contribution with fewer than six degrees of freedom, it is recommended that k be set equal to the two-tailed value of the Student's t for the number of degrees of freedom associated with that contribution and for the level of confidence required…*" (EURACHEM / CITAC, 2000).

2 MU FROM PT DATA; SINGLE RESULTS.

2.1 MU in the absence of Bias.

We have already discussed how equations 4 and 5 describe uncertainty estimation in the absence of bias and how this scenario reflects the current situation regarding uncertainty evaluation in AAR geochronology. Whilst it is perhaps, a little optimistic to expect bias to be completely absent from AAR data, it is not beyond the realms of possibility for bias and its associated uncertainty to be so small so as to be negligible. From the results of the proficiency study given in Part 1, clearly this is not actually the case, and bias would appear to play an important part in the accuracy of D/L values, even if at present it cannot be routinely determined by laboratories. Nonetheless, if hypothetically, bias was effectively negligible, the uncertainty of an individual laboratory's result would simply be the equivalent of their intermediate precision (s_{Rw}), determined during in-house method validation, for that particular matrix / analyte / concentration combination. However, this information was not provided by participants of the study.

In many instances, reported AAR uncertainty estimates are simply expressed as the standard deviation of results. Where these results were determined in a single run, an estimate of the run to run variability needs to also be included in any uncertainty estimate, otherwise uncertainty values will only represent the repeatability component and will most likely underestimate overall precision. In many cases, uncertainty estimates for AAR data are incompletely expressed either as the standard deviation of repeated measurements or the CV% equivalent.

In situations, where the only data available is a set of replicate values, and the analytical result is determined as the mean of these replicates, any standard uncertainty estimate would be more correctly reported as the standard uncertainty (or standard deviation of the mean, or standard error) (EURACHEM / CITAC, 2000, JCGM 100:, 2008), i.e.; $u(\varepsilon) = s_\varepsilon{}^2/n$ see section 1.3.1.

To demonstrate how measurement uncertainty could be calculated assuming replicate results were the only information available, standard uncertainties were calculated from the proficiency test results, for each participant for all amino acids in every test material. Standard deviations (*s*) for each participant's replicate data, were used to calculate the standard uncertainty of the mean in each case; i.e.;$u(\bar{x}) = s\sqrt{n}$, where *n* is the number of replicates. Relative standard uncertainties were then determined as; $RSU\% = ((u(\bar{x})) / \bar{x}) \times 100$.

To illustrate the effect of the expanded uncertainty on participants' results, expanded relative uncertainties were calculated as; $U\% = RSU\% \times k$, using coverage factors $k=2$ and $k=t_{(0.05,df)}$ for comparison, where the degrees of freedom were based on the number of replicate values in each case. These were then used to derive upper and lower confidence limits for participant's results. Full details and charts, are given in the reports for each amino acid in each test material, available at www.neaar.co.uk. As an example, Figure 5 illustrates this for alanine data in Mollusc (B) test material. 5(a) shows replicate values submitted by participants for rpHPLC, and as the mean and standard deviation for GC data. Comparison of charts

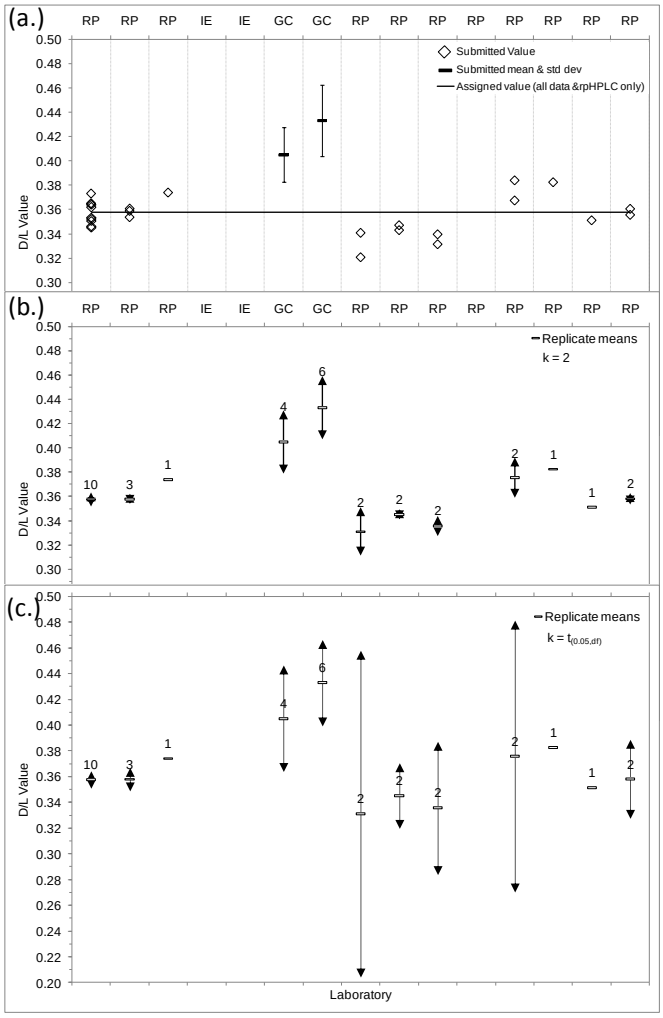


Figure 4; Effect of Expanded uncertainty on Participants' mean Alanine D/L values in Mollusc(B) test material. (a) Shows individual laboratories' rpHPLC replicate D/L values or GC D/L means and standard deviations determined, (b) Laboratories' mean D/L values and expanded confidence limits using *k=2* (c) Laboratories' mean D/L values and expanded confidence limits using $k=t_{(0.05,df)}$.

5(b) using a coverage factor $k=2$, with 5(c) using $k=t_{(0.05,df)}$ clearly highlights the problem of using small data sets; the smaller the value of n, the larger the t-value and the bigger the expanded uncertainty. It will always be a trade-off between quality or precision and laboratory resources both financially and in machine and analyst time. In this case, the greatest gain will be had by increasing the number of replicates from 2 to 3 see Figure 4(b). Subsequent replicate analyses will still make reductions in the expanded uncertainty but at a diminishing rate.

Regardless, repeatability is perhaps one of the least variable of the uncertainty components associated with a measurement result, and conceivably, one of the smallest. In practice, additional repetitions of analysis at this level is likely to have little if any influence on the final combined standard uncertainty estimate and it is far better to focus limited resources at higher level uncertainty components to give better precision estimates, such as increasing the number of site samples, where greater variability might be expected.

In routine analysis however, it would be completely unreasonable to expect a laboratory to carry out a full precision analysis, including run to run and day to day variability (let alone a bias evaluation) with every batch of samples. This is one reason why estimates of intermediate precision or intra-laboratory reproducibility derived during method validation are so useful. Providing that an in run check on repeatability is in agreement with validation data, intermediate precision estimates can be used with each batch of samples, whose matrices fall within the scope of the validated method. Intermediate precision represents a typical uncertainty estimate that could be achieved on any particular day, with any instrument, analyst or batch of reagents etc, depending on the reproducibility conditions employed.

Note that depending on the specified conditions, additional uncertainty contributions may be required to reflect uncertainty due to sampling, which is often outside the control of a laboratory technician. Uncertainty due to sampling is beyond the scope of the current paper and readers are directed to the joint Eurachem/EUROLAB/CITAC/Nordtest/AMC guidance document (Ramsey and Ellison, 2007). In the current context, imprecision associated with heterogeneity of the test materials might be considered a sampling uncertainty. However it is proposed that any variability between the individual test materials will be reflected in the uncertainty of the assigned value or between-laboratory precision estimates subsequently derived, and need not be counted separately.

Intermediate precision is only a part of the combined uncertainty estimate. This now needs to be combined with bias uncertainty contributions. The evaluation of *bias* from proficiency test data was the subject of Part 1 of this paper. The next section will now look in more detail at the evaluation of *uncertainty due to bias* and how this can be used with precision data to provide an overall combined standard uncertainty, u_c.

2.2 MU using Precision and Bias Components

Bias (*bias*) and its associated uncertainty (*u(bias)*) are often evaluated as part of a laboratory's method validation process by analysis of a certified reference material (CRM) or from spiking and recovery experiments. Bias uncertainty, together with the determination of intermediate precision estimates (or intra-laboratory reproducibility standard deviation (S_{Rw}), also derived from method validation precision experiments, can define the overall combined uncertainty for a measurement system (u_C), as already shown by equations 8a and 8b. This is referred to as the 'top-down' approach to measurement uncertainty determination (Barwick and Ellison, 2000b).

Where method validation data is available, performance in a proficiency test can provide verification of a laboratory's own uncertainty estimates, which should be comparable with the spread of their own PT results over time. However in the absence of reference materials or validation data, PT results can provide a valuable indication of the combined method and laboratory bias in routine analysis in its own right, which together with an estimate of the laboratory's intermediate precision, (S_{Rw}), can provide a value for the combined standard uncertainty (u_c).

It should be recognised that due to the bias uncertainties associated with a PT, any combined uncertainty estimate is likely to be larger than that resulting from the analysis of a certified reference material (CRM) by an individual laboratory. It is recommended that long term bias trends are observed to lessen the impact from a single proficiency test result and at least 6 rounds of testing are used to evaluate bias estimates (Magnusson et al., 2004)

In addition, it is recommended that intra-laboratory precision estimates (S_{Rw}) are determined from replicate analyzes of samples under *reproducibility* conditions over an extended period of time to take account of between run and general day to day variability. To simply use the repeatability standard deviation from replicate results submitted for the proficiency test is not a realistic representation of the overall precision and may contribute to smaller expanded uncertainties than might be otherwise be expected.

It is widely recognised that evaluation of PT data can be a valuable addition to the determination of measurement uncertainty. However, there is very little information provided by the main guidance documents on exactly how this should be done (JCGM 100:, 2008, EURACHEM / CITAC, 2000). In the Eurachem guide, it is suggested that the standard deviation of the normalized differences should be used. However the method adopted here has been derived from two main sources; the Nordtest Report TR 537[1] (Magnusson et al., 2004) that was produced as a handbook for the Nordic environmental testing laboratories and Eurolab's Technical reports[2] Nos 1/2006 and 1/2007 (EUROLAB, 2006, EUROLAB, 2007). All documents are freely downloadable and recommended for further reading on the subject.

The information thus presented should perhaps be considered more as an information exercise than a definitive measure of uncertainty. This is due to a number of reasons; such as the relatively small number of submitted results, uncalibrated data, and the potentially empirical nature of the methods, all of which increase the uncertainty of the assigned values. Also because of the absence of true intra-laboratory precision estimates (S_{Rw}) and the fact that not all laboratories supplied analytical replicate values. Nonetheless, the

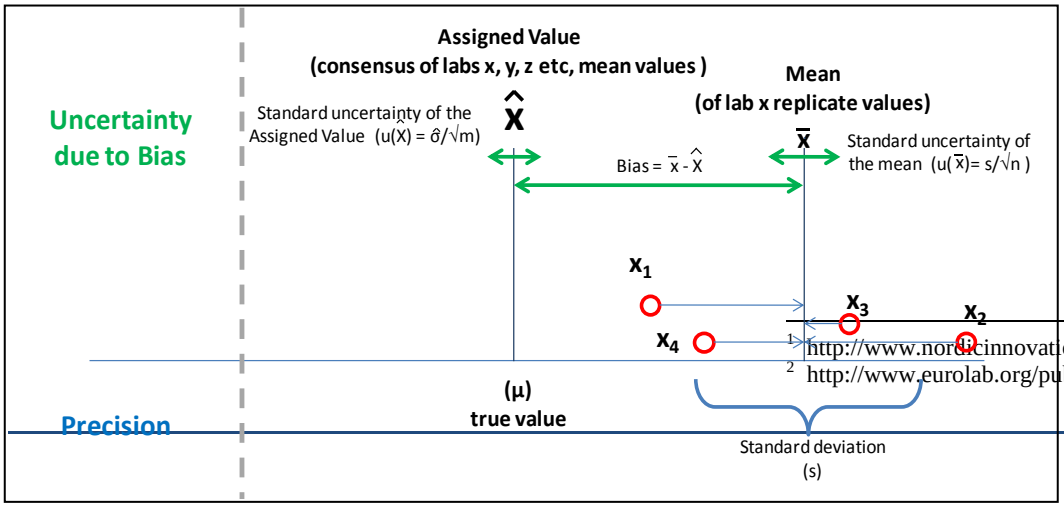


Figure 6; Bias and Precision Components derived from Proficiency test data, used in Measurement Uncertainty Estimation.

1 http://www.nordicinnovation.net/nordtestfiler/tec537.pdf
2 http://www.eurolab.org/pub/i_pub.html

data presented demonstrates how it can be possible to determine measurement uncertainty using proficiency test data which provide some interesting indicative values.

For those readers unfamiliar with measurement uncertainty estimation, distinguishing the various uncertainty components can be somewhat baffling. Figure 6 below helps to illustrate the sources and relevance of the different contributions due to precision and particularly those elements due to bias.

2.2.1 Standard uncertainty due to Bias ($u(bias)$).

The simplest expression for the bias uncertainty (*u(bias)*) is the experimental uncertainty of the laboratory mean $u(\bar{x})$ **plus** the uncertainty of the assigned value $u(\hat{X})$ where $u = s/\sqrt{n}$. Note; if a CRM was used as the test material, $u(\hat{X})$ can be taken from the specifications directly.

$$u(bias) = \sqrt{u(\bar{x})^2 + u(\hat{X})^2} = \sqrt{\frac{s_{\bar{x}}{}^2}{n_{\bar{x}}} + \frac{\hat{\sigma}^{\,2}}{m_{\hat{X}}}}$$

Where $s_{\bar{x}}$ = standard deviation of a laboratory's submitted result, $n_{\bar{x}}$ = number of laboratory replicates, $\hat{\sigma}$ = standard deviation of the assigned value, and $m_{\hat{X}}$ = number of laboratories' results contributing to the assigned value.

In routine analysis, bias should be accounted for and corrected for significant systematic effects. However in circumstances where this is not done by convention and the method is said to be empirical, any significant uncorrected bias should contribute to the combined uncertainty budget. Bias is determined as $bias = (\bar{x} - \hat{X})$ or as a relative value; $\frac{bias}{\hat{X}} = \left(\frac{\bar{x}-\hat{X}}{\hat{X}}\right)$ Where $\bar{x}$ = laboratory result (or the mean of replicate values) and $\hat{X}$ = the assigned value.

To determine whether the observed bias is significant or not, the *t* statistic is calculated and compared to the 2-tailed critical value for *n-1* degrees of freedom. If *t* is greater than or equal to the critical value, t_{crit} , then the bias is significant and an additional term to account for uncorrected bias in the result needs to be included in the combined uncertainty estimate (EURACHEM / CITAC, 2000).

t is calculated as;

$$t = \frac{1-Rec}{u(Rec)}$$

where ; $Rec = \bar{x}/\hat{X}$ and usually represents the recovery associated with the analysis of a CRM, and $u(Rec)$ is the same as *u(bias)* given above.

If $t \geq t_{crit}$, *Rec* is significantly different from 1 and the result $\bar{x}$ remains uncorrected, a bias correction term needs to be included in the combined uncertainty estimate.

However, this scenario is to some extent academic as the uncertainty of the assigned value in a proficiency test is likely to be much larger than that of a CRM (if one were available) and it is recommended to include the bias contribution in the uncertainty evaluation at all times regardless of whether $t \geq t_{crit}$ or not (Magnusson et al., 2004).

Thus, the bias uncertainty now becomes;

$$u(bias) = \sqrt{(\bar{x} - \hat{X})^2 + \frac{s_{\bar{x}}{}^2}{n_{\bar{x}}} + \frac{\hat{\sigma}^{\,2}}{m_{\hat{X}}}} \qquad or$$

$$u(bias) = \sqrt{(bias)^2 + u(\bar{x})^2 + u\left(\hat{X}\right)^2}$$

The combined uncertainty is now calculated as;

$$u_C = \sqrt{S_{Rw}^2 + u(\bar{x})^2 + u(\hat{X})^2 + (bias)^2} \qquad ()$$

Ideally, each laboratory's own amino acid / matrix specific intermediate precision estimate should be used. However this information wasn't available so each laboratory's own standard deviation taken from submitted results was used as a measure of precision, expressed as the CV. All components were calculated as the relative or normalized values, and expressed as percentages. Combined uncertainty estimates, u_c, were derived and Expanded uncertainties, *U*, for each participant were determined using a coverage factor *k=2*. This was to simplify the calculations whilst considering uncertainty components from various sources but also in order to enable direct comparability between laboratories and across analytes.

2.2.2 Results and Discussion.

Full details of these evaluations can be found in the individual reports from this study. However, Figure 7 (a-d:i), provide examples of histograms used to demonstrate the relative contributions to uncertainty from the various precision and bias components for each participant's results. For laboratories who did not provide replicate values, precision estimates could not be determined and so CV% contributions, shown as the black bars, are not present. Where both precision and bias components are present however, the combined standard uncertainty for each laboratory is shown with a cross. Figure 7 (a-d:ii), then illustrate the effect that expanding the combined standard uncertainties have on the mean of laboratories' replicate values, using a coverage factor *k=2* in all cases. The arrows on the error bars indicate the extent of the upper and lower confidence interval equivalent to 2 standard deviations. Laboratory values shown without error bars in the (ii) figures, are those who provided only a single D/L result (n=1).

Figure 7 (a) gives isoleucine (D-Allo/L-Ile) results in Standard Solution test material. As perhaps might be expected, all uncertainty contributions are very small. Relative bias results given here have been calculated using the assigned value derived from ALL participants' results, that is rpHPLC, HPLC-IE and GC data taken together. Thus the higher relative biases observed for the GC are due to a median based predominantly on rpHPLC. Whether this is the true value for this amino acid in this test material is not known, it's just the best estimate from the available information. GC data would perhaps suggest the true value is a little higher, but this cannot be confirmed without additional GC data.

Because there have been no preparation steps involved other than rehydration of the sample vial, what little variation there is between the rpHPLC data will be due to laboratory bias, instrumental effects and the rehydration stage, rather than actual method or matrix error influences. More significant bias effects are observed for the GC values and highlight method effects, although as previously mentioned, which one is in error is a matter of debate. For isoleucine in standard solution (D/L approximately 0.5), the combined uncertainty (u_c) range for GC data is between 3.6-10.5%, (Expanded uncertainty values will be double this, i.e. 7.2-21%). For reverse-phase data in the same test material, the combined uncertainty ranges are much smaller, between 0.71-2.5% for all participants.

Figure 7 (b) show data, again for isoleucine, but in ostrich egg shell test material; OES (A). Figure 7 (b.i) clearly shows the effect on standard uncertainty estimates due to sample preparation (method plus laboratory error contributions) and matrix effects, when compared to the standard solution data discussed previously. In this case, there is very little difference observed between GC or rpHPLC

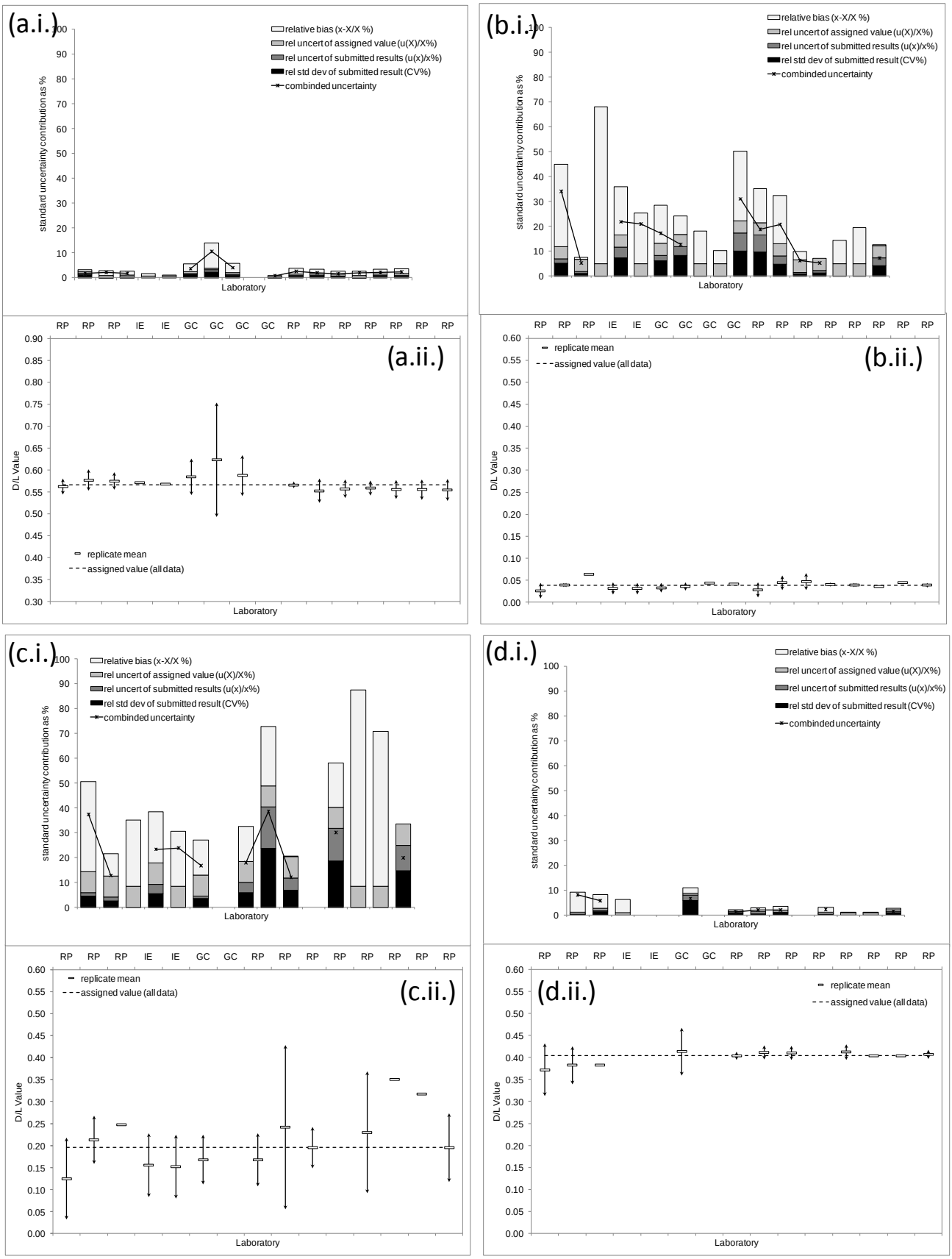


Figure 7; Precision and Bias Standard uncertainty contributions, their combined and expanded effect on submitted results. Figure (a) is for A/I in Standard Solution test material, (b) is A/I in OES(A) test material, (c) is A/I in Mollusc (B) test material and (d) is for aspartic acid in Mollusc (B) test material. Figures (i) demonstrate the relative contributions from the precision and different uncertainty sources. Figures (ii) show the effect on the participants' submitted replicate means, of expanding (k=2) these combined standard uncertainties, shown against the assigned value for the analyte.

data because the level of agreement both within and between laboratories is far more variable leading to larger (im)precision estimates and biases, masking the smaller ones. In the majority of cases, the biggest contribution is due to the actual bias component. For OES (A), the combined uncertainty range is between 5.2-34%. Thus it can be seen that if the analytical bias can be corrected, then the overall uncertainty estimate could be greatly improved. The combined and expanded uncertainties have been calculated as relative percentages, therefore, when translated into numerical confidence limits, the extent of the confidence interval is entirely dependent on the mean D/L value applied. For ostrich egg shell test materials the amino acid mean D/Ls are generally very low. So although the overall relative uncertainties may be fairly wide, the effect on the mean D/L value is minimal as shown in Figure 7 (b.ii). In contrast to this, Figure 7 (c.i & ii) show uncertainty data once again for isoleucine, but this time in Mollusc (B) test material. From the histogram (c.i), individual uncertainty components, expressed as relative percentages, are of a similar scale to those for OES (A), if perhaps a little larger. The combined uncertainty range for isoleucine in Moll(B) is 12-39%. However, because the mean D/L value for most of the amino acids in Mollusc (B), are much higher, the effect on the expanded confidence limits is far more noticeable. So, although in relative terms, in some situations uncertainty estimates may be similar, however in practice this translates to wider numerical uncertainty estimates for samples with higher D/L values, i.e.; those which are older or have a warmer temperature history.

The last two charts in Figure 7, (d.i & ii), show uncertainty data this time for aspartic acid but again in Mollusc (B) test material. These figures help to demonstrate the differences observed between different amino acids within the same matrix. (d.i) shows that the within and between laboratory agreement for aspartic acid is generally excellent and considerably better than the isoleucine looked at so far. Even in a complex matrix, where there will be method, laboratory and matrix effects all contributing to the uncertainty estimates, the combined uncertainty range is between 1.2-8.1%. This time, when these relative values are applied to an even higher mean D/L value, the effect on the confidence limits is not nearly so severe (Figure 7 (d.ii)).

So not only is uncertainty matrix and concentration dependent, it is also highly variable between different amino acids too.

3 MU FROM PT DATA; COMBINING MULTIPLE RESULTS

3.1 MU in the absence of Bias

Whilst it can be informative to visualize the uncertainty contributions from different sources, the greatest gain is to be had from observing bias performance over time.

For a "*well-behaving*" (performance is within the satisfactory range) single laboratory who has participated in several different rounds of proficiency testing for the same or similar analyte in the same or similar matrix, there will be no overall bias ($bias$). Where the uncertainty of the assigned values, $u(\hat{X})$ and the uncertainty of replicate values, $u(\bar{x})$ are small compared to the intermediate precision estimate s_{Rw}, determined through a laboratory's own method validation, then the standard uncertainty derived from results falling within the satisfactory range, could be said to be equivalent to the target standard deviation, σ_p, used for the assessment of the data, i.e.; because results comply with the target precision estimate in the absence of bias.

However, in this report, no values for target standard deviation, σ_p, have been given. Under these circumstances and assuming the absence of bias described above still holds for results falling within the satisfactory range, the uncertainty of a specific laboratory's submitted results would simply be equivalent to that laboratory's own intermediate precision or intra-laboratory reproducibility standard deviation s_{Rw}, etc., as previously discussed in section 2.1.

Because the absence of bias is an ideal scenario, and unrealistic, uncorrected bias and the uncertainty due to bias should always be included in the uncertainty budget, even if their overall contribution is small, at least until the analyst is confident that analytical results are free from systematic effects.

3.2 MU using Precision & Bias Root Mean Square

The procedure requires the calculation of the bias root-mean-square (RMS_{bias}%,) expressed as a percentage. When multiple results are being combined (i.e.; from a series of proficiency tests or in our case from a set of laboratories), the contribution due to bias ($bias$) and the uncertainty due to bias ($u(\bar{x})$), (i.e.; the standard uncertainty of the mean), can be directly replaced by the bias RMS_{bias}%. (EUROLAB, 2007). Note that the uncertainty of the reference material or assigned value still needs to be included in the overall calculation of the total uncertainty due to bias (*u(bias)*).

$$u(bias) = \sqrt{RMS_{bias}{}^2 + u(\hat{X})^2} \qquad \text{and}$$

$$RMS_{bias} = \sqrt{\sum(bias_i)^2 / p}$$

Where p is the number of proficiency test results being combined, $u(\hat{X}) = u(X_{ref})$ from equation 7 , and $u(\hat{X}) = \bar{s}_{\hat{X}} / \sqrt{\bar{m}_{\hat{X}}}$

The average standard deviation for the assigned values, $\bar{s}_{\hat{X}}$, and the average number of participants across all the tests, $\bar{m}_{\hat{X}}$, can be determined and used to calculate an average uncertainty $u(\hat{X})$ for the tests.

From equation…. the combined uncertainty (u_c) now becomes;

$$u_C = \sqrt{s_{Rw}^2 + RMS_{bias}{}^2 + u(\hat{X})^2}$$

3.2.1 Amino Acid uncertainty estimates.

In order to evaluate the effect of uncertainty on an individual laboratory's series of PT results, an estimate of that laboratory's intermediate precision is first required. However this information was not provided by participants. An estimate of the average standard uncertainty of the mean of replicate values could be used but as already discussed, this would be an underestimation of that laboratory's overall precision and is therefore not ideal.

Therefore rather than evaluating a series of results for a single laboratory, a series of results for a single amino acid have been used. Thus, for every amino acid in each test material, the combined uncertainty has been determined using all of the participants' submitted results for that particular amino acid in a given matrix.

The s_{Rw}, has been replaced by the standard deviation of the assigned value ($\hat{\sigma}$), (i.e.; sMAD), as the precision estimate, $u(\hat{X})$ is simply the uncertainty of the assigned value for that amino acid and the RMS_{bias} is derived using all the PT results for each amino acid in the test material. Results have also been evaluated for rpHPLC subset, due to a separate assigned value being available. Following this, the Expanded uncertainty (U) can be derived using an appropriate coverage factor, such as k=2, as has been used in these examples.

3.2.2 Results and Discussion.

Table 1 presents the results of this evaluation. All values are represented as the relative percentage. This data provides an

Table 1: Standard uncertainty contributions, plus the combined and expanded uncertainties, for every amino acid in each test material, averaged across all submitted results.

amino acid	Opercula Test Material					OES(A)					OES(B)				
	$\hat{\sigma}$ as RSD%	$u(\bar{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2	$\hat{\sigma}$ as RSD%	$u(\bar{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2	$\hat{\sigma}$ as RSD%	$u(\bar{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2
Asx D/L-all[a]	1.02	0.28	4.97	5.08	**10.16**	3.84	0.99	6.96	8.01	**16.02**	4.14	1.07	12.12	12.85	**25.71**
Asx D/L-rpHPLC	1.17	0.35	1.70	2.09	**4.19**	3.76	1.13	3.14	5.03	**10.06**	3.47	1.05	5.84	6.87	**13.74**
Glx D/L-all[a]	1.47	0.41	8.82	8.95	**17.90**	8.32	2.15	9.69	12.95	**25.90**	6.85	1.90	12.98	14.80	**29.60**
Glx D/L-rpHPLC	1.29	0.39	6.41	6.55	**13.10**	12.72	3.83	9.45	16.30	**32.61**	5.88	1.77	13.91	15.20	**30.41**
Ser D/L-rpHPLC	1.41	0.43	1.39	2.03	**4.05**	1.27	0.38	3.35	3.60	**7.20**	3.57	1.08	2.27	4.36	**8.72**
Arg D/L-rpHPLC	21.76	7.25	22.2	31.92	**63.83**	11.55	3.85	15.6	19.79	**39.58**	7.10	2.37	7.49	10.59	**21.18**
Ala D/L-all[a]	5.15	1.43	4.25	6.83	**13.66**	12.25	3.27	10.99	6.78	**33.56**	16.09	4.15	13.88	21.65	**43.30**
Ala D/L-rpHPLC	5.14	1.55	3.58	6.45	**12.90**	10.24	3.09	7.12	12.85	**25.71**	8.34	2.51	7.9	11.76	**23.52**
Val D/L-all[a]	6.99	1.94	8.25	10.99	**21.98**	14.13	3.65	16.46	22.00	**43.99**	11.49	2.97	18.12	21.66	**43.32**
Val D/L-rpHPLC	7.58	2.28	5.5	9.64	**19.27**	13.23	3.99	9.22	16.61	**33.22**	2.14	0.64	8.74	9.02	**18.04**
Phe D/L-all[a]	2.87	0.79	4.94	5.77	**11.53**	5.14	1.33	9.05	10.49	**20.98**	9.49	2.74	7.26	12.26	**24.52**
Phe D/L-rpHPLC	3.01	0.91	4.54	5.52	**11.04**	3.12	0.94	4.38	5.46	**10.92**	5.51	1.66	7.91	9.78	**19.56**
D-Aile/L-Ile-all[b]	35.21	9.09	26.48	44.99	**89.97**	20.37	4.94	21.45	29.99	**59.98**	14.82	3.59	17.95	23.55	**47.11**
D-Aile/L-Ile -rpHPLC	16.94	5.11	20.44	27.03	**54.07**	18.92	5.71	24.1	31.17	**62.33**	7.28	2.19	19.05	20.51	**41.02**
Leu D/L-all[a]	16.12	5.10	16.21	23.42	**46.84**	18.81	5.22	18.87	27.15	**54.30**	10.18	2.94	11.84	15.89	**31.78**
Leu D/L-rpHPLC	7.86	2.78	12.9	15.36	**30.72**	8.24	2.75	15.32	17.61	**35.22**	8.36	2.96	8.43	12.24	**24.48**
Tyr D/L-rpHPLC	1.99	0.89	4.78	5.25	**10.50**	6.89	2.61	6.48	9.81	**19.63**	10.20	3.86	5.32	12.13	**24.27**

[a] = rpHPLC and GC data [b] = rpHPLC, GC and HPLC-IE data

Error! Reference source not found. **(continued).**

amino acid	Standard solution Test Material					Mollusc(A)					Mollusc(B)				
	$\hat{\sigma}$ as RSD%	$u(\hat{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2	$\hat{\sigma}$ as RSD%	$u(\hat{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2	$\hat{\sigma}$ as RSD%	$u(\hat{X})$ as RSU%	RMS_{Bias} %	Combined/Expanded u_c%	U% k=2
Asx D/L-all[a]	2.02	0.54	13.58	13.74	**27.48**	2.24	0.65	3.05	3.84	**7.68**	2.58	0.78	3.55	4.46	**8.92**
Asx D/L-rpHPLC	0.48	0.15	1.35	1.44	**2.88**	2.24	0.68	3.12	3.90	**7.80**	2.55	0.81	3.72	4.58	**9.17**
Glx D/L-all[a]	2.94	0.79	7.84	8.41	**16.82**	2.57	0.74	7.58	8.04	**16.08**	6.00	1.81	9.15	11.09	**22.18**
Glx D/L-rpHPLC	1.50	0.47	1.54	2.20	**4.40**	2.39	0.72	6.89	7.33	**14.66**	4.16	1.32	10.08	10.98	**21.97**
Ser D/L-rpHPLC	2.38	0.75	1.47	2.89	**5.79**	4.38	1.32	11.89	12.74	**25.48**	5.60	1.77	11.44	12.86	**25.72**
Arg D/L-rpHPLC	4.76	1.68	11.09	12.19	**24.37**	22.93	8.11	21.93	32.75	**65.50**	20.49	7.24	22.13	31.02	**62.04**
Ala D/L-all[a]	6.12	1.64	12.62	14.12	**28.24**	9.52	2.75	14.34	17.43	**34.87**	6.98	2.01	8.24	10.99	**21.97**
Ala D/L-rpHPLC	1.28	0.40	12.42	12.49	**24.98**	9.35	2.82	7.35	12.22	**24.45**	5.97	1.89	4.50	7.71	**15.41**
Val D/L-all[a]	1.90	0.53	7.36	7.62	**15.24**	11.45	3.30	11.89	16.83	**33.67**	10.06	3.03	12.74	16.52	**33.03**
Val D/L-rpHPLC	1.49	0.47	8.18	8.33	**16.65**	11.85	3.57	10.57	16.27	**32.55**	11.01	3.48	10.43	15.56	**31.11**
Phe D/L-all[a]	1.53	0.41	1.77	2.38	**4.76**	8.27	2.39	9.84	13.07	**26.14**	2.51	0.76	8.52	8.92	**17.83**
Phe D/L-rpHPLC	0.85	0.27	1.02	1.36	**2.71**	7.04	2.12	8.67	11.37	**22.74**	1.91	0.60	8.34	8.58	**17.16**
D-Aile/L-Ile-all[b]	2.69	0.70	3.15	4.20	**8.40**	24.75	6.62	39.55	47.12	**94.25**	30.69	8.51	33.33	46.10	**92.19**
D-Aile/L-Ile -rpHPLC	1.01	0.32	1.57	1.89	**3.79**	27.56	8.31	34.3	44.78	**89.56**	17.51	5.54	32.75	37.55	**75.09**
Leu D/L-all[a]	3.71	1.12	5.6	6.81	**13.62**	13.60	4.30	15.42	21.01	**42.01**	15.47	5.47	13.59	21.31	**42.61**
Leu D/L-rpHPLC	1.13	0.40	2.41	2.69	**5.38**	17.25	6.10	13.85	22.95	**45.89**	12.44	5.08	15.21	20.29	**40.59**
Tyr D/L-rpHPLC	-	-	-	-	**-**	0.57	0.26	7.03	7.06	**14.12**	3.10	1.55	3.47	4.90	**9.80**

[a] = rpHPLC and GC data [b] = rpHPLC, GC and HPLC-IE data

indication of typical uncertainty estimates which a laboratory might be expected to achieve for each amino acid. It provides a convenient comparison between the different matrices for any given amino acid, and assists in the comparison of overall uncertainty for different amino acids in the same test material.

Table 1 presents the combined and expanded uncertainty values for amino acids in different test materials and demonstrates how different uncertainty contributions need to be considered. For each amino acid in each test material, data have been evaluated using all submitted results and also separately for rpHPLC data only, where separate assigned value uncertainty estimates were also determined. In nearly every case, the uncertainty estimates were made worse by the inclusion of all methods, as might be expected, therefore the following account relates only to the rpHPLC data, unless otherwise stated, so as to provide a single method comparison. Figure 8 helps to summarize this data and shows the rpHPLC expanded uncertainty values ranked in order of increasing uncertainty; position 1 being the most precise, or the least variable data.

Figure 8(a) presents data for Opercula and Standard Solution test materials. In the Standard Solution test material (D/L ≈ 0.5), in the majority of cases, the expanded uncertainty values (*U%*) ranged between 2.7-5.8%, with phenylalanine and aspartic acid giving the lowest values of less than 3%. Uncharacteristically for isoleucine, the reverse-phase data was remarkably well-behaved. It was no surprise to see that arginine was amongst the poorest performers however it was unexpected to find that alanine gave the widest uncertainty overall with U=25%, and valine not far behind with U=16.7%. For some time arginine has been suspected of having stability issues which may be responsible for the variability in observed data, but these results would also strongly suggest that alanine and possibly valine may be experiencing similar complications when in solution.

For bleached opercula test material, this time serine and aspartic acid are the best performers with expanded uncertainties hovering just above 4%. Tyrosine, phenylalanine, alanine, and glutamic acid then group together with estimates of U ranging between 10.5-13%, followed by valine at 19.3%, leucine 31%, and finally isoleucine and arginine both giving estimates over 50%.

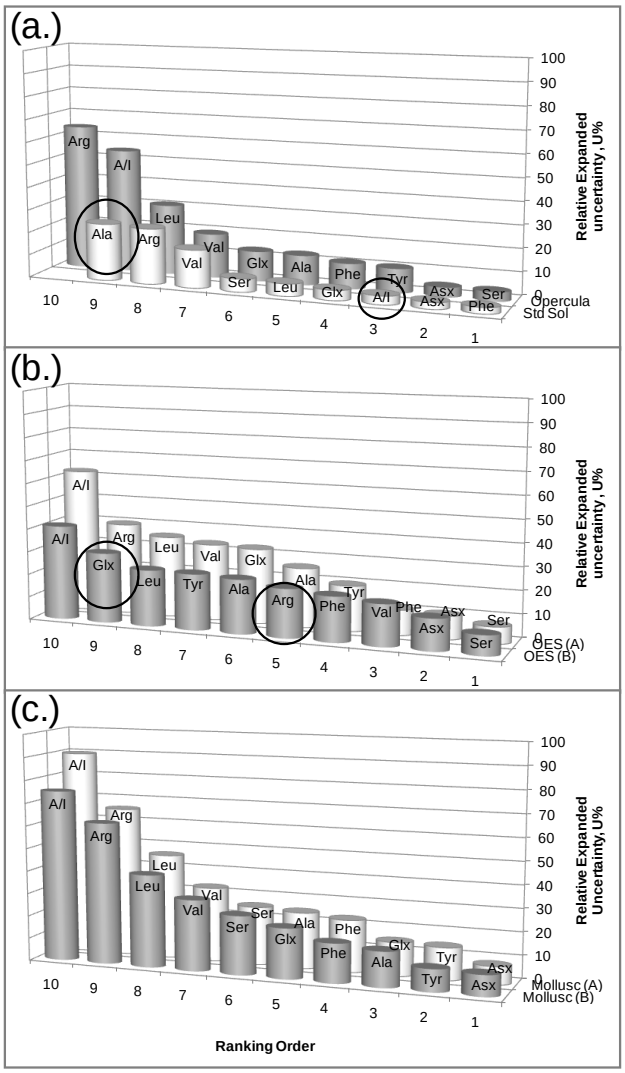
Figure 8(b) compares the ostrich egg shell test materials (OES(A) and OES(B)). In both test materials, serine had the lowest uncertainty, in both cases being less than 10% with aspartic acid not too far behind with expanded uncertainties of 10.1% and 13.7% respectively. In the bleached OES(A), phenylalanine's expanded uncertainty was close to aspartic acid with U = 11%. Following this came tyrosine (U=20%) and alanine (U=26%) and then glutamic acid, valine and leucine grouped at around 33-35%. Arginine then second to last with 40% and isoleucine lastly with U greater than 60%. By comparison, the majority of amino acids in OES(B), were in fairly close agreement with each other, ranging from 18-24.5% (valine, phenylalanine, unusually arginine, alanine, tyrosine and leucine). However, glutamic acid was particularly variable (U=30.4%) by comparison to the other amino acids in unbleached OES, which appeared to improve on bleaching by moving the Glx position higher up the ranking in OES(A). Finally once again the widest uncertainty estimate in OES(B) was that of isoleucine with 41%. It is interesting to observe the similarity in ordering between the bleached Opercula amino acids and those of the bleached OES(A), with the exception of tyrosine and phenylalanine with swap over.

The last chart is Figure 8(c) and compares bleached Mollusc shell (A) test material with unbleached Mollusc shell (B). In both cases, the ordering of amino acids is almost identical, with only glutamic acid and alanine swapping positions, alanine showing less agreement in bleached mollusk shell and glutamic acid showing better. For both test materials, aspartic acid have the smallest expanded uncertainties; OES(A), 10% and OES(B), 13.7%. The next best amino acid was tyrosine followed by the others with steadily increasing uncertainty estimates, to end with arginine with over 60% and finally isoleucine with 90% in the case of Mollusc (A) or 75% for Mollusc (B).

To summarize therefore, whilst there are some unusual exceptions observed in the data, it would appear that in the majority of cases, aspartic acid and serine provide the closest agreeing data; that is the smallest overall uncertainties, whilst isoleucine is generally the most variable amino acid by rpHPLC, closely followed by arginine and then leucine. The remaining amino acids; glutamic acid, alanine, valine, phenylalanine, and tyrosine then order themselves inbetween, depending on the matrix and treatment carried out but also on the concentration of the amino acid, remembering all the time that these are relative uncertainty estimates.

Ironically, the bleaching carried out in order to isolate a closed system of amino acids, does not appear to significantly improve the uncertainty estimates, either in terms of the precision, (i.e.; the

Figure 8; Ranked Amino Acids based on rpHPLC derived Expanded Uncertainties (*U%*) in the six different Test Materials. (a) Plots Opercula and Standard Solution test material data. Note the unusually tight isoleucine data (A/I) in position 3 in Standard Solution, and the unusually wide alanine data in position 9. Note also that tyrosine was not present in the Standard solution. (b) compares the bleached test material OES (A) with the unbleached OES (B). Note the "well-behaved" arginine data and wide glutamic acid in OES (B). (c) then compares the bleached Mollusc (A) test material with the unbleached Mollusc (B).



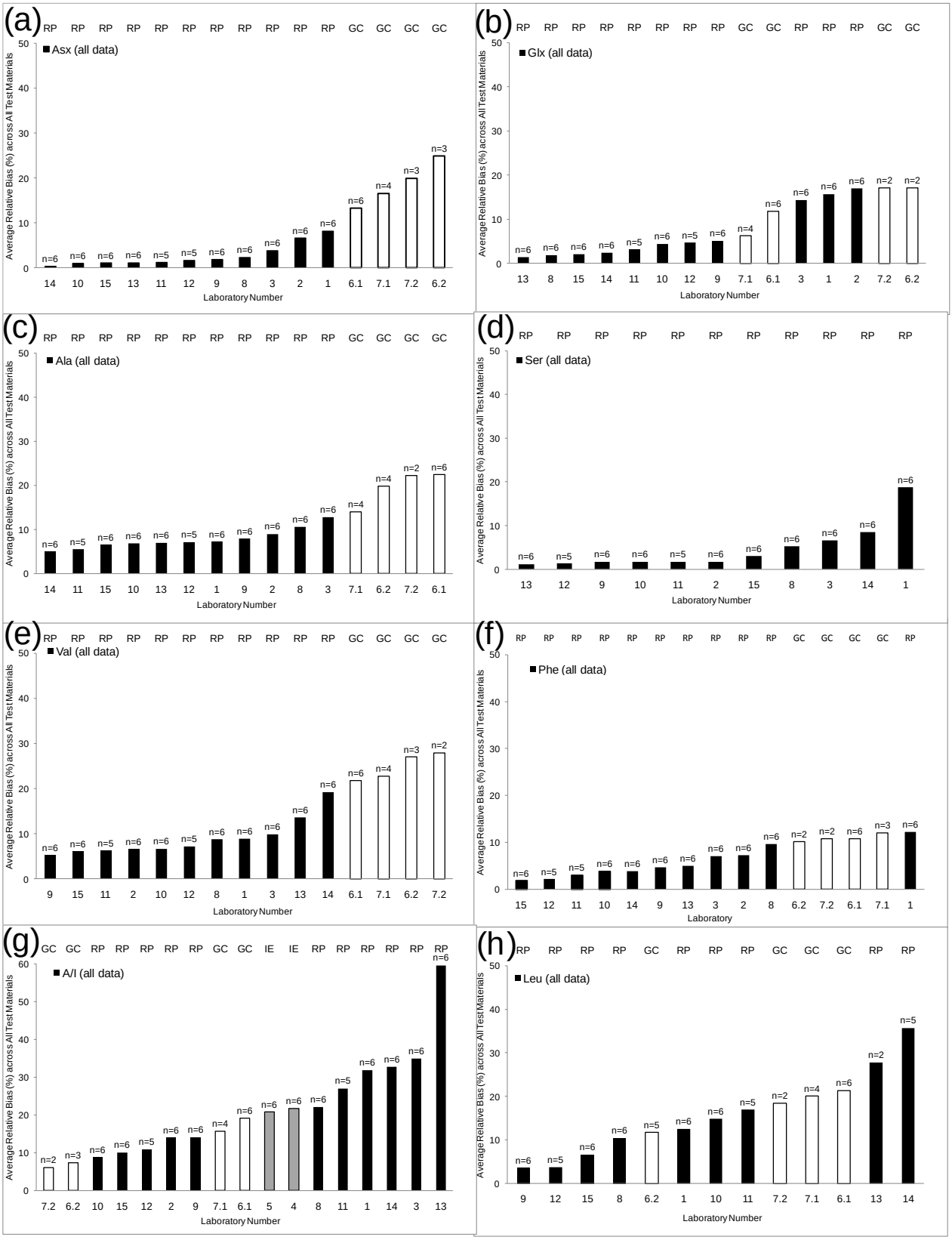


Figure 9; RMS_{Bias}% Histograms showing individual participant's standard uncertainties due to analytical bias, evaluated across all six test materials. (a) aspartic acid, (b) glutamic acid, (c) alanine, (d) serine, (e) valine, (f) phenylalanine, (g) isoleucine, (h) leucine

standard deviation of the assigned value; sMAD) or in terms of the observed bias as seen in Table 1. Close examination of the tabulated data shows that bleaching will reduce uncertainty estimates for some amino acids, expand uncertainty estimates for others and occasionally make little if any difference at all.

Table 1 presents if you like, a worst case scenario, reflecting both inter-laboratory and method error effects. Strictly speaking, the uncertainty of the assigned value, $u(\widehat{\boldsymbol{X}})$, is unnecessary as it is already incorporated in to the standard deviation of the assigned value $\hat{\sigma}$ (or sMAD), used as our precision estimate. However, the contribution is generally small by comparison and not likely to significantly influence the combined uncertainty estimate, but it's inclusion serves as a reminder to include reference value uncertainties From the data it can be seen, that generally speaking, it is the analytical bias (RMS%) that contributes the largest component to the combined uncertainty estimate and the uncertainty of the assigned value or the uncertainty of the reference value that is generally the smallest. Control of bias is paramount in producing accurate and reliable data. It can therefore be appreciated how correction of significant bias could reduce this element substantially, and reduce overall uncertainty estimates for each amino acid.

3.2.3 Individual Laboratory RMS_{Bias} estimates.

Although, for reasons already discussed, it is not possible to fully evaluate individual laboratory's uncertainties, we can however determine and compare the ***RMS_{bias}%* component**, using their own results across all six test materials. Each laboratory's RMS_{bias}% results have been calculated and are plotted as histograms in Figure 9. It is recognised that there will be some differences in biases observed between different matrices, for example, biases in standard solution tend to be much smaller than those in mollusc shell. However, these effects will be common to all the participants who submitted results for all six test materials, and is still a valid basis for comparison. The only complication arises for those participants where $p \neq 6$, and a more matrix specific comparison is recommended from the specific PT reports.

3.2.4 Results and Discussion

Individual laboratory's RMS% have been evaluated for every amino acid, combining data from each of the six test materials. These results have been arranged in an ordered manner and are presented as histograms in Figure 9.

In spite of the small numbers of participants providing GC and IE data, there is a strong indication from the data presented in Figure 9, that for aspartic acid, alanine, valine, and possibly phenylalanine (with the exception of a single high rpHPLC laboratory) GC data quantify slightly higher than the rpHPLC results, and even for glutamic acid, it would appear that GC results, 6.2 and 7.2 are also consistently high. Interestingly GC results 6.1 and 7.1 were determined using peak area data, whilst 6.2 and 7.2 used peak height data. Knowing this it can also be seen that GC peak height D/L values appear to quantify higher than the peak area data for aspartic acid, glutamic acid and valine. By comparison, for isoleucine, GC peak height data quantify much lower than everything else, whilst the GC peak area values and HPLC-IE fall centrally together. Whilst it is possible that once again we are observing empirical differences it is not clear whether these differences are significant or simply that they fall towards the upper end (or lower end) of normal distributions. If D/L quantification was not method dependent, then it would be expected that results of both rpHPLC, GC and IE would be randomly distributed. Both leucine and isoleucine data are more suggestive of this. Interestingly, for isoleucine, both GC peak area data and HPLC-IE data, cluster together, and demonstrate the close agreement between these methods, on which the development of the technology has been based.

4 MU FROM PT DATA; ANOVA (RSD_R%)

4.1 Bias as a random variable

So far, the methods looked at for determining measurement uncertainty from proficiency test data, have considered precision and bias as independent components, which together give an overall estimate of the combined standard uncertainty. However, Section 1.3.2 has already discussed how individual laboratory biases, when viewed from a higher perspective, can be seen as a random variable describing between laboratory precision. It is this unique relationship between precision and bias which is utilized in the assessment of inter-laboratory collaborative trial data using a one-way ANOVA, (analysis of variance) to determine method precision estimates (Horwitz, 1995, AOAC, 2000, RSC Analytical Methods Committee, 1995). ANOVA allows us to evaluate independent sources of uncertainty in a single step, separating the random error components from the laboratory effect.

ANOVA is used to determine the within laboratory or repeatability precision (s_r) and the between laboratory precision (s_L) which together provide an estimate of the overall precision for a given method, the reproducibility standard deviation (s_R).

Within laboratory precision is expressed as the repeatability standard deviation, s_r, and represents an inter-laboratory approximation of random error effects. Often the s_r is more conveniently represented as the relative repeatability standard deviation and expressed as a percent, (RSD_r%).

s_R is the reproducibility standard deviation and a measure of the **overall precision for an analyte** in the specified test material. Again, this is more conveniently expressed as the relative standard deviation of reproducibility, (RSD_R%). s_R incorporates both the within laboratory random error effects and the between laboratory bias, the later being conveniently expressed as a precision estimate, and is a single measure of the (im)precision or uncertainty of the measurement procedure.

Conventionally, between laboratory precision estimates, s_L, tend not to be used on their own. Uncorrected bias will exaggerate the between laboratory variance component, resulting in wider RSD_R% values.

Because a collaborative trail is method prescriptive, it is assumed that method bias and its uncertainty are zero. In principle the same technique could be used as a novel approach to determining overall uncertainty from proficiency test data. If ANOVA was applied to participants' replicate data from a proficiency study, incorporating all the method variability, the reproducibility values would now reflect the additional uncertainty due to routine method differences too; i.e., one rung higher up the ladder of errors.

ANOVA calculations were carried out for all the amino acids in each test material, allowing for unequal numbers of replicates (Miller and Miller, 2005).

All submitted results were included in this evaluation without removal of outliers as would otherwise be done with collaborative trail data. On this occasion it was the intention to observe the behavior of all submitted results rather than to define best practice. Because GC data were reported as the mean and standard deviation, individual replicate data were not available and so ANOVA has not been carried out on any GC data. However all rpHPLC amino acids have been assessed and HPLC-IE results have also been evaluated separately for isoleucine, albeit with limited data. Table 2 therefore summarizes method specific repeatability and reproducibility precision estimates for each amino acid. The mean D/L values have been determined as the average from all individual replicate values and are likely to be more sensitive to extreme values and vary slightly from the assigned values determined as the median of submitted results.

Table 2: Summary of Inter-Laboratory Precision Estimates (Repeatability & Reproducibility) determined using a One-Way ANOVA from Participants' submitted replicate results.

amino acid	Opercula Test Material					OES(A)					OES(B)				
	p	N	[1]mean D/L	RSD_r%	RSD_R%	p	N	[1]mean D/L	RSD_r%	RSD_R%	p	N	[1]mean D/L	RSD_r%	RSD_R%
Asx D/L-rpHPLC	11	29	0.564	0.54	2.43	11	26	0.364	1.04	4.22	11	26	0.210	0.40	8.16
Glx D/L-rpHPLC	11	29	0.157	0.52	6.23	11	25	0.085	4.83	10.95	11	26	0.056	0.47	15.46
Ser D/L-rpHPLC	11	29	0.656	1.12	1.77	11	27	0.329	0.70	2.69	11	26	0.111	0.82	3.05
Arg D/L-rpHPLC	9	17	0.776	19.53	27.95	9	15	0.139	3.96	14.09	9	15	0.101	3.75	8.68
Ala D/L-rpHPLC	11	29	0.268	2.01	4.72	11	27	0.094	3.29	7.10	11	26	0.063	10.21	11.27
Val D/L-rpHPLC	11	29	0.135	4.33	6.93	11	27	0.029	7.61	12.22	11	27	0.019	4.72	11.05
Phe D/L-rpHPLC	11	29	0.306	5.18	5.90	11	27	0.077	4.40	6.37	11	26	0.053	2.35	8.29
D-Aile/L-Ile -rpHPLC	11	29	0.194	13.65	34.51	11	27	0.035	5.04	29.65	11	27	0.026	4.21	29.94
D-Aile/L-Ile -HPLC-IE	2	4	0.137	2.45	2.90	2	5	0.031	6.04	6.04	2	4	0.024	0.00	0.00
Leu D/L-rpHPLC	8	24	0.289	8.11	14.60	9	24	0.063	4.81	18.52	8	23	0.050	11.16	12.80
Tyr D/L-rpHPLC	5	10	0.273	2.24	5.42	7	11	0.078	1.07	7.18	7	11	0.059	3.44	7.49
amino acid	Standard solution Test Material					Mollusc(A)					Mollusc(B)				
	p	N	[1]mean D/L	RSD_r%	RSD_R%	p	N	[1]mean D/L	RSD_r%	RSD_R%	p	N	[1]mean D/L	RSD_r%	RSD_R%
Asx D/L-rpHPLC	10	23	0.499	0.84	1.60	11	28	0.412	0.71	4.63	10	26	0.390	0.53	4.93
Glx D/L-rpHPLC	10	23	0.553	0.67	1.70	11	28	0.214	1.70	9.94	10	26	0.188	1.11	15.56
Ser D/L-rpHPLC	10	23	0.402	0.97	1.71	11	28	0.490	2.58	22.98	10	26	0.395	2.37	18.04
Arg D/L-rpHPLC	8	16	0.376	3.24	11.49	8	14	0.659	24.41	27.07	8	14	0.650	24.10	25.73
Ala D/L-rpHPLC	10	23	0.489	0.69	12.31	11	28	0.420	2.86	6.54	10	26	0.356	2.45	4.38
Val D/L-rpHPLC	10	23	0.438	0.43	6.53	11	28	0.198	5.56	10.27	10	26	0.167	5.79	10.16
Phe D/L-rpHPLC	10	23	0.492	0.75	1.17	11	28	0.266	8.57	13.86	10	26	0.235	8.57	16.01
D-Aile/L-Ile -rpHPLC	10	23	0.561	0.75	1.55	11	28	0.233	11.42	36.65	10	26	0.187	10.84	37.61
D-Aile/L-Ile -HPLC-IE	2	4	0.576	0.11	0.29	2	4	0.186	0.81	4.41	2	4	0.154	3.91	3.91
Leu D/L-rpHPLC	8	18	0.597	1.76	3.21	8	23	0.312	13.35	16.40	6	20	0.240	13.69	16.44
Tyr D/L-rpHPLC	-	-	-	-	-	5	9	0.239	2.33	6.11	4	8	0.218	1.89	4.20

p = no of sets of results N = total no of replicate values [1] = mean of the participants' individual replicate D/L values

4.2 Results and Discussion

Because this assessment has been carried out using participants' results for individual amino acids in each test material, RSD_R% values presented in Table 2, will be directly comparable to the amino specific combined standard uncertainties given previously in Table 1. Potentially they may also be comparable to the inter-laboratory CV% from Part 1 Table 2 too, although the CV% is perhaps closer to the between laboratory precision (s_L), rather than being a true representation of the overall uncertainty.

However, the relative repeatability standard deviations, RSD_r%, should also be comparable to the average intra-laboratory CV% given in Part 1; Table 2, as both measurements are reflections of the imprecision due to random error effects only.

With only a few exceptions, generally, it would appear that both the RSD_r% and RSD_R% values quantify slightly higher than the intra and inter-laboratory CV% equivalents (Part 1; Table 2), whilst RSD_R% values quantify slightly lower than the combined uncertainty estimates from Table 1. However, all three precision values (i.e.; inter-lab CV%, RSD_R% and u_c as RSU%), for any particular amino acid were of a similar scale, suggesting that whilst ANOVA carried out this way may be a novel approach for evaluating the combined uncertainty, it would appear not to be an unreasonable one, in many instances lying midway between the other two approaches.

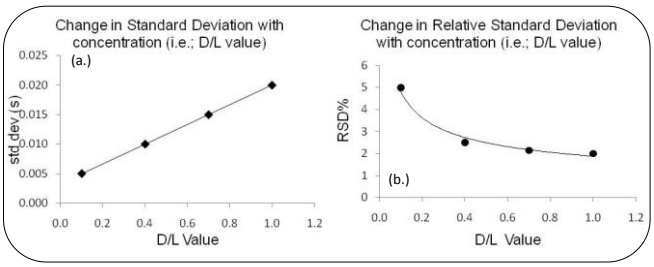
4.3 Predicting Reproducibility

In many instances the precision of chemical analysis is often dependent on concentration, i.e.; the observed standard deviation increases as analyte concentrations increase. Figure 10 illustrates this relationship using some theoretical data. In this example, it can be seen that the standard deviation increases with D/L value, but decreases when expressed as a relative value. It can be seen that initially this decrease is steep (since even a small difference at a low concentration can have a big effect), and then plateaus out. William Horwitz was the first to report on this relationship between precision and concentration (Horwitz, 1982). It was found that the reproducibility standard deviation values obtained from collaborative trial data for a particular group of analytes, varied in a predictable manner with analyte concentration. This relationship, referred to as the Horwitz curve or the Horwitz equation, has been found to be widely applicable to many different analytes and is used widely in other sectors for predicting target values for standard deviation by proficiency test providers and in quality control.

The Horwitz equation requires concentration to be expressed as a mass fraction, and so is not universally applicable. However, it was considered that an evaluation of the amino acid RSD_R values relative to the D/L values, may be insightful, possibly indicating similar relationships within our own test materials, in spite of the additional variability due to method differences in the current study.

For each amino acid, mean D/L values from all six test materials were plotted on the same chart, against their respective precision estimate. Figure 11 (a), (b) and (c) presents precision data for isoleucine, aspartic acid and valine respectively, as examples. Figures (i) and (ii) use reproducibility data determined using ANOVA; Figures (i) shows how the standard deviation of reproducibility (s_R) changes with D/L value, whilst Figures (ii) shows how the relative standard deviation of reproducibility (RSD_R%) (data from Table 2) relates to D/L. For comparative purposes, GC and IEx inter-laboratory precision data for ILC-A, B and C originally presented by Wehmiller (1984) and expressed as (inter-laboratory) CV% are plotted in Figures (iii). Lastly, results of an ANOVA carried out on the PT test material homogeneity data as a measure of intra-laboratory repeatability, i.e.; in the absence of inter-laboratory bias, are given in Figures (iv).

Figure 10 Theoretical relationship between analyte level, standard deviation (s) and relative standard deviation (RSD%). 10(a) shows increasing standard deviation with increasing analyte level, 10(b) shows how the standard deviation decreases relative to the analyte level, so at low levels, the standard deviation actually becomes more significant.



Initially, plotted data appeared confusing, especially for some amino acids, which showed no obvious correlations. However, on further inspection, relationships between the predominantly calcitic matrices, (opercula and the two ostrich egg shell test materials), became clearly visible in all amino acids. Further, in many instances, trend lines could be drawn between these three data points and the Standard Solution test material, which gave curves characteristic of those suggested in Figure 10. Figure 11(a, ii.), isoleucine RSD_R% is the exception to this as it is known to exhibit poor reproducibility by rpHPLC. However, isoleucine by GC or IEx (Figure 6(a, iii)) and intra-lab isoleucine by rpHPLC (Figure 6(a, iv) agree with expectations.

Mollusc shell behavior however, was not quite so obvious and in most cases was set apart from the calcite data. Submitted results for the two mollusc shell test materials were the most widely variable of all the six test materials, the matrix clearly being more challenging. Mollusc shell data will therefore reflect these larger precision estimates, which in turn are likely to interfere with any predictable patterns in this matrix. Wehmiller's GC data clearly indicate there are correlations within the three ILC mollusc shell materials used in the original trial, but these are not directly super-imposable over the calcitic curves.

The results for the ANOVA on the homogeneity D/L data (Figures (iv)) act as a control and allow us to observe the behavior of the data under more controlled repeatability conditions, i.e.; without the influence of method or laboratory bias affecting the data. Overall intra-laboratory precision values thus determined are much smaller than their inter-laboratory reproducibility equivalent values and give an indication of the improvement that might be expected once method and laboratory bias have been brought under tighter control. However there remains insufficient data to clarify the behavior of mollusc shell matrix.

These differences in predictive precision between the calcitic matrices (opercula and ostrich egg shell) and the calcitic/arogonitic mollusc shell, suggest differences in position and functionality of amino acids in the biomineral proteins and their racemization tendency. It should be recognised that the data used in this analysis includes additional variability due to method differences in the proficiency study and that far tighter estimates are likely to be achievable if data from a method specific interlaboratory collaborative trial were available. Species specific differences in the D/L value have been reported for opercula (refs……) and it would be an interesting extension to see whether these differences are carried across to the precision estimates or whether in fact, reproducibility is predictable across species sharing a common biomineral matrix.

5 CONCLUSIONS

In Part 2 of this paper we have introduced the concept of a hierarchy of accumulating error contributions, known as the "ladder of errors" and considered how precision and sources of bias contribute to uncertainty determination. The most common approaches to measurement uncertainty evaluation have been

summarized with particular emphasis on the "Top-down" approaches and how different methods account for different levels of uncertainty, eg; intra-laboratory method validation will evaluate precision and single laboratory + method bias but doesn't account for inter-laboratory bias, a collaborative trail will reflect inter-laboratory bias but doesn't include method bias etc. Proficiency test data, quite uniquely has the potential to account for all the levels of error, if evaluated across participants results reflecting typical variations in routine methods.

Precision and bias estimates obtained from proficiency test data, presented in Part 1, have been used to derive combined and expanded uncertainty estimates. Examples for calculating uncertainty related to a single proficiency test result have been given in section 2, whilst section 3 demonstrated how it could be possible to derive a representative uncertainty estimate across a series of results. To illustrate how this could be done, analyte specific combined uncertainty estimates were determined, using the RMS_{Bias} approach on laboratories' submitted results. In practice this technique would usually be applied by individual laboratories to monitor uncertainty across a series of proficiency tests. However, to do this requires knowledge about the individual laboratory's intermediate precision estimate from in-house precision studies, usually done as part of the method validation, alternatively the target value for standard deviation used to assess data in the proficiency test could be used for laboratories whose results lay within the satisfactory range. As neither of these pieces of information were available it was not possible to determine combined uncertainty estimates for each laboratory. However, it was possible to determine each individual laboratory's RMS_{Bias} contribution using their own results for each of the six test materials. These were presented as ordered histograms to show how the uncertainty contribution due to bias affected different laboratories for the various amino acids.

Finally, a novel approach to evaluating analyte uncertainty using ANOVA was presented. This approach is more commonly applied to collaborative trial data and evaluates bias contributions as a random variable. Where participants in a proficiency test have provided replicate results, the reproducibility standard deviation RSD_R, can provide an estimate of the overall standard uncertainty. When assessing PT data derived using different methods, this can provide a unique way of incorporating the elusive method bias into an uncertainty estimate, representing one of the highest rungs of the "ladder of errors". This could conceivably be extended to include matrix and even concentration effects in the same way. However, laboratory accreditation and method validation will usually specify method, matrix, analyte and even concentration range, which generally avoids the need to include these higher level errors in uncertainty estimates.

It can be seen from the data presented in Table 2, that, with the exclusion of arginine, isoleucine and leucine, (known to be analytically problematic by rpHPLC), the analytical or instrumental repeatability standard deviation is better than 1% for all other amino acids in Standard Solution, increasing to perhaps 1.5% when expanded to 2 standard deviations. The overall uncertainty as RSD_R%, ranges from 1.2 – 12.3% (1 std dev). In comparison, the matrix bound amino acids have much wider uncertainties reflecting method bias (from preparation and extraction stages) as well as matrix specific effects. Interestingly, the ostrich egg shell whose submitted results were the tightest of all the sets of data, gave relative uncertainty estimates of a similar scale to those of the mollusc shell, whose distribution of submitted results was the most variable. This will, at least in part, be because of the low level of D/L values in the ostrich egg shell. D and L concentrations will be heading towards the limit of quantification for the method and the instrument's resolution capability. Thus even a small variation in D/L value at a low level will have an unusually large effect. ANOVA repeatability uncertainty estimates for amino acids in biomineral matrices range from <1% up to 12%, and for reproducibility uncertainty, from about 2% up to 23%.

Inspite of these estimates having been evaluated across laboratories carrying out the same or very similar rpHPLC methods, these results and those of the performance assessments carried out in Part 1, suggest there is still considerable variability observed between reported values for the same test material. The inability to correct for bias in routine analysis due to the absence of reference materials for calibration and quality control is a serious issue which needs urgent attention. Accurate uncertainty determination could be achieved through a method prescriptive collaborative trial, where method bias is effectively zero. Only then will the true level of variability, that any laboratory could reasonably expect to achieve for a given amino acid in a specific matrix, be determined. Remaining test materials could then be used as fit-for-purpose reference materials with reference values derived from the consensus of the laboratory results and repeatability and reproducibility uncertainty values having been determined.

True measurement uncertainty evaluation can be a very uncomfortable issue for any analyst who naturally wants reliably tight data. Instrumental precision estimates, can give very small standard deviations and coefficients of variation, but these represent only one of the smallest and lowest contributions to uncertainty on the "ladder of errors".

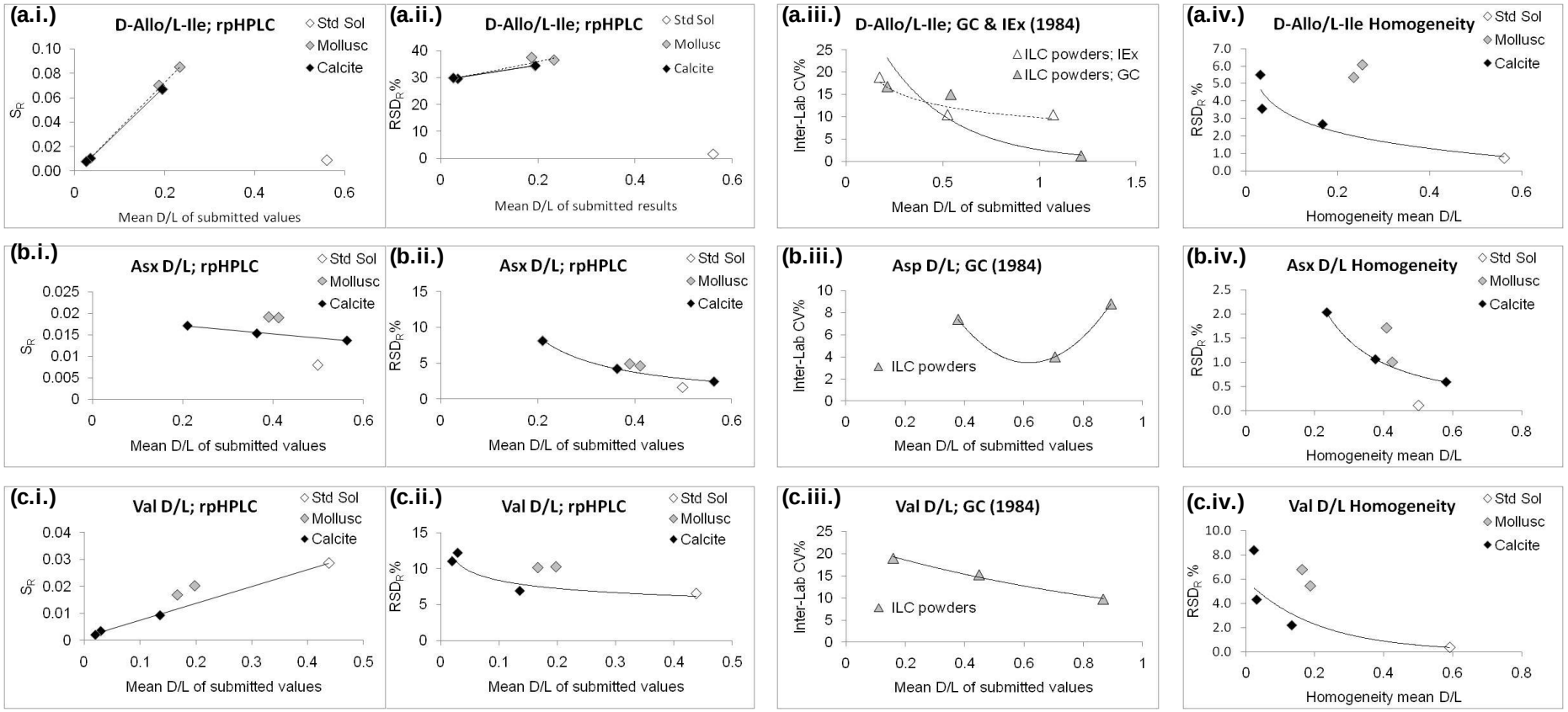


Figure 11; Examples of the Relationship between D/L value and precision. 6(a) relates to isoleucine data, 6(b) relates to aspartic acid data and 6(c) relates to valine data. Figures a, b & c (i) and (ii) illustrate how D/L values relate to *ANOVA inter-laboratory reproducibility* precision estimates; (i) shows the reproducibility standard deviation (s_R) and (ii) gives the relative reproducibility standard deviation (RSD_R%). Figures a, b & c (iii) relate D/L with original GC and IEx *inter-laboratory precision* estimates for ILC-A powder data (Wehmiller 1984) and Figures a, b & c (iv) present the *intra-laboratory* relative reproducibility standard deviation (RSD_R%) or *intermediate* relative precision derived using ANOVA on the PT

6 REFERENCES

AMC (1995) Uncertainty of Measurement: Implications of its use in Analytical Science. *The Analyst,* 120**,** 2303-2308.

AOAC (2000) AOAC Official Methods Program Manual Part 12. Appendix D: Guidelines for Collaborative Study Procedures to Validate Characteristics of a Method of Analysis. *Available from; http://www.aoac.org/vmeth/omamanual/omamanual.htm*. AOAC International.

BARWICK, V. J. & ELLISON, S. L. R. (2000a) Development and Harmonisation of Measurement Uncertainty Principles Part (d): Protocol for uncertainty evaluation from validation data. *VAM Technical Report,* LGC/VAM/1998/088.

BARWICK, V. J. & ELLISON, S. L. R. (2000b) The evaluation of measurement uncertainty from method validation studies. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 5**,** 47-53.

CURRELL, G. & DOWMAN, A. (2005) *Essential Mathematics and Statistics for Science,* Chichester, John Wiley & Sons Ltd.

EURACHEM / CITAC (2000) Guide CG 4: Quantifying Uncertainty in Analytical Measurements. 2 ed., Available from; http://www.citac.cc/QUAM2000-1.pdf.

EUROLAB (2006) Technical Report No. 1/2006. Guide to the evaluation of measurement uncertainty for Quantitative test results. *Available from; http://www.eurolab.org/docs/technical%20report/EL_11_01_06_387%20Technical%20report%20-%20Guide_Measurement_uncertainty.pdf.*

EUROLAB (2007) Technical Report No. 1/2007. Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation. *Available from; http://www.eurolab.org/pub/i_pub.html.*

HORWITZ, W. (1982) Evaluation of analytical methods used for regulation of foods and drugs. . *Analytical Chemistry,* 54**,** 67A-76A.

HORWITZ, W. (1995) IUPAC Protocol for the design, conduct and interpretation of method-performance studies.

ISO 5725 (1994) Accuracy (trueness and presicion) of measurement methods and results - Part 2; Basic method for the determination of repeatability and reproducibility of a standard measurement method., International Standards Organisation.

ISO 21748 (2010) Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation. International Standards Organisation.

JCGM 100: (2008) Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). 1 ed., Available from; http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

JCGM 200: (2008) International Vocabulary of Metrology - Basic and general concepts and associated terms (VIM). Available from; http://www.bipm.org/en/publications/guides/vim.html

MAGNUSSON, B., NAYKKI, T., HOVIND, H. & KRYSELL, M. (2004) NORDTEST Report TR 537. Handbook for calculation of measurement uncertainty in Environmental Laboratories. *Available from; http://www.nordicinnovation.net/nordtestfiler/tec537.pdf*. 2 ed.

MILLER, J. N. & MILLER, J. C. (2005) *Statistics and Chemometrics for Analytical Chemistry,* Harlow, England., Pearson Education Ltd.

RAMSEY, M. H. & ELLISON, S. L. R., (EDS.), (2007) Eurachem/EUROLAB/CITAC/Nordtest/AMC Guide; Measurement uncertainty arising from sampling. A guide to methods and approaches., Available from the Eurachem secretariat.

RSC ANALYTICAL METHODS COMMITTEE (1995) Uncertainty of measurement: implication of its use in analytical science. *The Analyst,* 120**,** 2303-2308.

THOMPSON, M. (2000) Towards a unified model of errors in analytical measurement. *The Analyst,* 125**,** 2020-2025.

WESTAWAY, R. (2009) Calibration of decomposition of serine to alanine in Bithynia opercula as a quantitative dating technique for Middle and Late Pleistocene sites in Britain. *Quaternary Geochronology,* 4**,** 241-250.