# D [illegible]

A M C , AMC B N 53

27 2012

DOI: 10.1039/ 2 90034

[illegible] O [illegible] B [illegible]

A recent meta-analysis[1] has reviewed available studies of reported uncertainties in inter-laboratory exercises and examined additional examples of metrology comparisons in analytical chemistry. Although the number of such studies is modest, all those reviewed show evidence that uncertainty is more often underestimated than overestimated – that is, differences among laboratories are usually greater than the reported uncertainties would suggest. As an example of this common occurrence, Fig. 1 shows the results, with their reported uncertainties, for lead in tuna, produced by participants in IMEP 20. (IMEP is the International Measurement Evaluation Programme organised by the European Institute for Reference Materials and Measurements (IRMM), Geel, Belgium). Fig. 2 shows the observed distribution of the sorted results, together with bootstrapped estimates of the expected distribution had the uncertainty estimates been correct. On the basis of the reported uncertainties, the between-laboratory standard deviation should be 0.031 ppm: the observed robust value (that is, outliers discounted) was 0.122 ppm.

This kind of occurrence is neither especially novel nor peculiar to analytical chemistry, as we can see from the classic 1972 paper by Youden[2] on estimates of the velocity of light. But why now, two decades after the publication of the Guide to the Expression of Uncertainty in measurement ("the GUM"),[3] should this still happen?

‘ differences among laboratories are [illegible]

## G M

The GUM provided three things. First, it provided some basic concepts, such as the concept of measurement uncertainty itself as a summary figure that includes all possible effects, random these 'standard uncertainties' should be combined using the established rules for combining variances; and the then quite radical idea that uncertainties for both random and systematic effects should be treated identically, no matter whether they were estimated from statistical analysis ("Type A") or from other sources ("Type B") such as calibration certificates, manufacturer specifications or professional judgement. Finally, the GUM provided a particular approach to the combination of uncertainties, based on an equation (the 'measurement model') that was assumed to include all known significant effects on the measurement result. This particular methodology has been described as the 'bottom up' approach because of its focus on building up an uncertainty budget from individual parts.

## B G M

Since the publication of the GUM, other approaches have become available that respect the same principles but use alternative combination methods or simpler models. In particular, the every individual contribution through to the use of a much simpler (if less informative) summary figure of performance. And indeed both approaches are widely used in practice. But does either of these extremes guarantee an accurately estimated uncertainty?

‘Dark uncertainty seems to be not [illegible]

## [illegible]

The measurement community often seems polarised towards one or other of two extreme points of view.

- The ‘bottom-uppers’ or ‘splitters’ believe that the deconstruction procedure should be exhaustive, continued to provide a complicated complete ‘model’ of the procedure. ‘Splitters’ assert (correctly in most instances) that reproducibility standard deviation tends to underestimate standard uncertainty because inter alia the effects of method bias are not accounted for. The issue of traceability is also raised: how is the outcome traceable to the SI?
- The ‘top-downers’ or ‘lumpers’ believe that deconstruction should be terminated at the earliest possible point that gives rise to a reasonable estimate of uncertainty. The extreme version of the ‘lumper’ approach is simply to use reproducibility standard deviation (obtained by replication of the entire procedure in different laboratories) as their estimate of standard uncertainty. ‘Lumpers’ take the view (again correctly in most instances) that analytical procedures involve chemical interactions so numerous and complex that it is usually impossible to build a comprehensive model. There are both hidden influences on the result and unknown interactions between overt influences. The outcome is ‘dark uncertainty’,[1] present in the result of the measurement but not visible in the uncertainty budget. However, all of the effects, known and unknown (but excluding method bias), will be taken into account in reproducibility precision, because each laboratory using the procedure will explore the variable space differently and more-or-less at random. Because of this, dark uncertainty will be manifest in the reproducibility standard deviation, even though we do not know its source.

Advocates of both of these views, then, claim that the alternative method tends to under-estimate uncertainty. But these contentions are open to testing. A recent study of chemical measurement[8] has found a strong tendency for reproducibility standard deviation to be greater than an estimate based on a splitter approach, by a factor of about 1.5–2. And reproducibility standard deviation itself is potentially too small: it does not account for method bias. Dark uncertainty seems to be not only ubiquitous but almost inevitable in chemical measurement. So what should the analyst do?

## [illegible]

An obvious place to start is to check whether uncertainty estimates are realistic. This is covered in another AMC Brief,[9] so we will not discuss it in detail here. But as a simple rule of thumb, an uncertainty estimate much better than typical reproducibility standard deviations $s_R$ found for relevant methods and test materials should be reviewed as suspect. Where no relevant studies are available, relevant guidance (often regulatory) on acceptable performance may be a useful guide. And in the food analysis sector, Horwitz’s compilations have demonstrated a strong general tendency for reproducibility standard deviation to be about twice the associated repeatability standard deviation $s_r$ (that is, $s_R \approx 2s_r$) so a general tendency for uncertainty estimates in a laboratory to be less than $2s_r$ should be regarded as suspect. Where we should look, once we have identified a potential problem, depends on the approach we have taken for our uncertainty estimate. The GUM assumes that we have an equation that describes, quantitatively, all known, significant effects on the result. This is one obvious place to look for missing uncertainties.

## [illegible]

In principle, we can apply the GUM approach to the equation in the pesticide example above. A cursory examination might suggest that chromatographic peak areas can be estimated with a (relative) standard uncertainty of about 1%, that masses and volumes can be determined with uncertainty near 0.1% and that the stock solution uncertainty (which depends on further weighings and volumetric operations) could be known with relative uncertainty well under 1%. Combining these in the usual way gives a relative standard uncertainty of the order of 1.5–2%. We might see a repeatability relative standard deviation of 5–15% on spiked test materials, so the estimated relative uncertainty could be, perhaps, 10%.

This may be a fair summary of the combination of known calibration uncertainties and observed repeatability – and indeed confirms very nicely that we need take no further care over our instrument and glassware calibration1(1p.60.7(which343thainti1-A()TJT(5 perhaps some additional allowances. The critical questions are then "which estimate of precision?" and "of which measurement?"

Precision can be estimated from any set of repeated observations, from re-presentation of an extract to an instrument, through repetition of the complete measurement with no changes in calibrations, operator or equipment, to repetition by different laboratories. But the estimates of precision we get under these different conditions are very different, and we need to choose the right one. In one study of uncertainties reported in proficiency tests, it was