Accred Qual Assur (2008) 13:83–89
DOI 10.1007/s00769-007-0330-1

GENERAL PAPER

A practical procedure for assigning a reference value and uncertainty in the frame of an interlaboratory comparison

Stefaan Pommé · Yana Spasova

Received: 16 July 2007 / Accepted: 7 November 2007 / Published online: 1 December 2007

Abstract A method is suggested for the calculation of a reference value and its uncertainty to be used in the frame of an interlaboratory comparison (ILC). It is assumed that the reference value of the measurand is determined independently from the ILC round. It is derived from a limited set of measurement results obtained from one or several expert laboratories. The procedure involves three stages: (1) check of the experimental data and possible corrections; (2) check of the consistency of data, and possibly increase of the uncertainties in order to attain internal consistency; (3) choice between fully, partially or un-weighted mean.

Keywords Interlaboratory comparison · Reference value · Mean · Uncertainty

Introduction

In the frame of an interlaboratory comparison (ILC), usually the organiser provides a reference value and corresponding uncertainty for the measurand. We consider the case where this value is based on measurements performed by one or more expert laboratories. The laboratories are expected to provide an unbiased result with a complete and realistic uncertainty budget. The organiser evaluates and combines the available data in a statistically appropriate manner.

S. Pommé (✉) · Y. Spasova
European Commission, Joint Research Centre,
Institute for Reference Materials and Measurements (IRMM),
Retieseweg 111, 2440 Geel, Belgium
e-mail: stefaan.pomme@ec.europa.eu

Y. Spasova
e-mail: yana.spasova@ec.europa.eu

International standards in the field of ILCs [1–3] leave room for interpretation on how a reference value and uncertainty should be calculated.

The ISO guide 43 [1] and ISO 13528 standard [2] provide five 'methods' and support alternative methods, "*provided that they have a sound statistical basis*".

The ISO guide 43 [1] states the following:

> *"The following statistics may be appropriate when assigned values are determined by consensus techniques:*
>
> i) *mean, which may be weighted or transformed (e.g. trimmed or geometric mean)*
> ii) *median, mode or other robust measure*."

The recommendation in the harmonised protocol for proficiency testing [3] is more restrictive:

> "*Even when uncertainty estimates are available, unweighted robust methods (i.e. methods taking no account of the individual uncertainties) should be used to obtain the consensus value and its uncertainty [...]*"

It is clear that the guides do not agree upon the use of the uncertainties provided by the expert laboratories. In our opinion, the laboratories should (try to) provide complete and realistic uncertainty budgets. Their use in the relative weighting of the data should depend on the degree of reliability that is reached in the uncertainty assessments.

In this work, the case is considered in which the reference value is calculated from a few data. It could also be applied in cases where there are many data, as an alternative to other robust measures like, e.g., the median. For the assignment of a reference value, a logical three-stage procedure is followed: (1) identification and correction of errors and unrealistic uncertainties; (2) detecting discrepancies and achieving consistency; (3) establishing the reference value and its uncertainty. A

similar structure has also been identified by others (see e.g. ref. [4]). The most particular feature of the proposed procedure in this work is the possibility to move smoothly between a weighted and an unweighted mean. A flowchart of the proposed procedure is supplied in the Appendix.

Stage 1: Identification and correction of errors and unrealistic uncertainties

As the coordinator of the ILC is responsible for assigning the reference value, one could argue that this includes checking the experimental data provided by the laboratories. In the first stage, he could scrutinise the data for possible errors and in particular, check whether the stated uncertainty budget is realistic. From experience, we know that uncertainties are often underestimated (see e.g. refs. [5, 6]), hence unrealistic values should be adapted. Possibly, this stage may involve re-evaluation and/or elimination of unreliable data, and/or initiation of additional experimental work.

Stage 2: Detecting discrepancies and achieving consistency

Now consider a set of measurement data with their uncertainty, as provided by the expert laboratories:

$$x_{\mathrm{i}} \pm u_{\mathrm{i}} \quad (i = 1, \ldots n)$$

in which n is the number of available data, x_{i} is the i^{th} measured value and u_{i} its standard uncertainty.

It is assumed that the common uncertainty components, e.g. related to instrument, method or basic physical constants, are negligible (or temporarily excluded from the budget).

An ideal data set would be internally consistent, i.e. the data scatter would not be larger than what can be expected from the declared uncertainties. This can, for example, be tested by calculating the Birge ratio ($\chi_n = R_{\mathrm{B}}$):

$$R_{\mathrm{B}} = \frac{s_{\mathrm{ext}}}{s_{\mathrm{int}}} \tag{1}$$

where:

- s_{ext} is the 'external' uncertainty:

$$s_{\mathrm{ext}}^2 = \frac{1}{n-1} \frac{\sum w_i (x_i - x_{\mathrm{w}})^2}{\sum w_i} \tag{2}$$

- s_{int} is the 'internal' uncertainty:

$$s_{\mathrm{int}}^2 = \left(\sum \frac{1}{u_i^2} \right)^{-1} \tag{3}$$

- x_{w} is the weighted mean:

$$x_{\mathrm{w}} = \frac{\sum x_i \cdot w_i}{\sum w_i} \tag{4}$$

- w_i is the weighting factor:

$$w_i = \frac{1}{u_i^2} \tag{5}$$

If $R_B \leq 1$, one can say that the data look consistent and move on to stage 3 of the procedure.

If $R_B \geq 1$, one is probably dealing with discrepant data. If the discrepancy is too big, one should consider looking for outliers, possible significant mistakes (cf. stage 1), eventually even decide to derive no reference value from them until better information becomes available. Even though the literature abounds in procedures to discern possible outliers (see e.g. [7] and references in [8]), they should be used sparingly as one can loose the 'correct' value and/or a clear indication of neglected uncertainty components [8].

If the discrepancy is rather mild, one may consider using the data after an appropriate increase of the uncertainties (see e.g. ref. [9] for different methods). A possible way of proceeding is to increase all reported uncertainties with a constant a until $R_{\mathrm{B}} \leq 1$:

$$u_i' = \sqrt{u_i^2 + a^2} \tag{6}$$

This way, at the end of stage 2, one has assured that the data set is internally consistent. There is of course no guarantee that all systematic uncertainty components are covered, since common uncertainty components do not appear from the data scatter. At least one can compensate for overly optimistic uncertainty claims and avoid the underestimation of the uncertainty of the reference value that would result from them.

One should also realise that sample inhomogeneity contributes to the possible discrepancy between expert laboratory results. Therefore, in principle, the uncertainty through inhomogeneity should be taken into account by the ILC organiser before calculating the Birge ratio. By adding this component to the uncertainties in stage 2, its propagation into the reference value is already taken into account and no further action is required in phase 3.

In this work, we define the uncertainty of the unweighted mean, $\overline{x} = \frac{\sum x_i}{n}$:

$$s_{\mathrm{u}} = \sqrt{\frac{1}{n} \left(\frac{\sum (x_i - \overline{x})^2}{n-1} \right)} \tag{7}$$

and the uncertainty of the 'adjusted' weighted mean:

$$s_{\mathrm{w}} = \sqrt{\left(\sum \frac{1}{u_i'^2} \right)^{-1}} \tag{8}$$

Stage 3: Establishing the reference value and its uncertainty

There is no definitive rule on how the reference value has to be derived from the data set. The median is a robust option if a sufficiently large data set is available, but the calculation of its uncertainty may pose a problem. The most commonly accepted approaches are a weighted or unweighted mean. The ideal case corresponds to the weighted mean of a consistent data set with correct uncertainties (see e.g. ref. [10]). In practice, the quality of the reported uncertainty is not always at a level where it can be fully trusted for relative weighting of data. In the extreme case that relative uncertainties are completely unrealistic, they should be disregarded and one should revert to an unweighted mean. We provide a formula that covers both extremes as well as intermediate cases:

- x_{ref} is the proposed reference value, calculated as a (partially weighted) mean:

$$x_{\mathrm{ref}} = \frac{\sum x_i \cdot w'_i}{\sum w'_i} \tag{9}$$

- $u(x_{\mathrm{ref}})$ is the corresponding uncertainty ($k = 1$):

$$u^2(x_{\mathrm{ref}}) = \left(\sum w'_i\right)^{-1} \tag{10}$$

- w'_i is the proposed weighting factor:

$$w'_i = \left[\left(\frac{u'_i}{S}\right)^{\alpha} S^2\right]^{-1} \tag{11}$$

for $0 \leq \alpha \leq 2$ and $u'_i = \sqrt{u_i^2 + a^2}$ and

- $S^2 = n \cdot \max(s_{\mathrm{w}}^2, s_{\mathrm{u}}^2)$ a measure for the variance per datum:

$$S^2 = n \cdot \max\left\{\left(\sum \frac{1}{u_i'^2}\right)^{-1}, \frac{1}{n}\left(\frac{\sum (x_i - \overline{x})^2}{n-1}\right)\right\} \tag{12}$$

One could consider expanding the external uncertainty by a correction factor due to the limited sample size. This is a known procedure for the unweighted mean of Gaussian distributed data, where the sample standard uncertainty is multiplied with an appropriate t-factor (cf. Student's t-distribution). It has not been implemented explicitly in our approach.

The presented equations for the reference value show similarity with the L_{p} estimator, which calculates a mean that smoothly varies between the sample median ($p = 1$), mean ($p = 2$) and mid-range ($p = \infty$) by continuously varying the power parameter p [11, 12]. However, the concepts are different, as the L_{p} estimator discards the available information on the uncertainty of the data, while the power parameter α in Eq. 11 controls the influence of the assigned uncertainty on the weighting factor.

One can easily recognise the special case of a fully weighted mean ($\alpha = 2$):

$$w'_i = \frac{1}{\left(\frac{u'_i}{S}\right)^2 S^2} = \frac{1}{u_i'^2} \tag{13}$$

leading indeed to the known uncertainty formula for a weighted mean:

$$u^2(x_{\mathrm{ref}}) = \left(\sum \frac{1}{u_i'^2}\right)^{-1} = s_{\mathrm{w}}^2 \tag{14}$$

By fulfilling the condition that $R_{\mathrm{B}} \leq 1$ in Eq. 1 (using w'_i instead of w_i), one makes sure that the uncertainty on the reference value cannot be smaller than what follows from the observed spread of the data.

Also the case of an unweighted mean is easily obtained ($\alpha = 0$):

$$w'_i = \frac{1}{\left(\frac{u'_i}{S}\right)^0 S^2} = \frac{1}{S^2} = \text{constant} \tag{15}$$

with the required uncertainty value

$$u^2(x_{\mathrm{ref}}) = \left(\sum \frac{1}{S^2}\right)^{-1} = \frac{S^2}{n} = \max(s_{\mathrm{w}}^2, s_{\mathrm{u}}^2) \tag{16}$$

In this case, by the definition of S, one avoids that the uncertainty goes below that of the weighted mean, i.e. what follows from the stated uncertainties.

Depending on the trust that one has in the uncertainties reported by the reference laboratories, one shall decide on full, partial or no weighting for calculating the reference value and its associated uncertainty. In some cases one will find an intermediate correlation between the deviations, $(x_i - x_{\mathrm{ref}})$, and the corresponding uncertainties, u'_i, and decide to use, e.g., $u_i'^{-1}$ as a relative weighting factor rather than $u_i'^{-2}$. This is achieved by applying $\alpha = 1$ in Eq. 11. Such an approach would be well-founded, for example, in the field of primary standardisation measurements of (radio)activity, as a systematic study of all available data in the Key Comparison DataBase (BIPM, Paris) shows that the deviations are rather proportional to $u^{0.5}$ than to u [5].

At this point, the uncertainty $u(x_{\mathrm{ref}})$ is complemented with the uncertainty components that have up to now been taken out of consideration, such as common uncertainty components and instability of the material [13]. The expanded uncertainty of the assigned value is obtained by multiplying the standard uncertainty $u(x_{\mathrm{ref}})$ by a coverage factor k, depending on the required level of confidence:

$$U(x_{\mathrm{ref}}) = k \cdot u(x_{\mathrm{ref}}) \tag{17}$$

Hypothetical example

Initial stage

Consider a hypothetical ILC that is supported by experimental data of the measurand by three expert laboratories (see Fig. 1a):

$x_1 \pm u_1 = 80 \pm 20$
$x_2 \pm u_2 = 108 \pm 1$
$x_3 \pm u_3 = 95 \pm 10$

Stage 1

Experts scrutinise the data and find the uncertainty u_2 unrealistically low, as the best attainable uncertainty is estimated to be $u_2 = 5$. The data set is adapted accordingly (Fig. 1b).

$x_1 \pm u_1 = 80 \pm 20$
$x_2 \pm u_2 = 108 \pm 5$
$x_3 \pm u_3 = 95 \pm 10$

Stage 2

The weighted mean and the internal and external uncertainties are calculated:

$x_w = 104.2$
$s_{ext} = 5.2$
$s_{int} = 4.4$
$s_u = 8.1$

The Birge ratio, $R_B = 1.2$, is slightly larger than one. Hence, we may have a discrepant data set. The data are scrutinised again, but one finds no apparent mistakes. Now the ILC coordinator decides whether to proceed with the exercise and to assign a reference value or not. He decides to add a constant a to the uncertainties in order to reduce R_B to 1 (see Fig. 1c).

$a = 6.5$
$x_1 \pm u_1 = 80 \pm 21$
$x_2 \pm u_2 = 108 \pm 8$
$x_3 \pm u_3 = 95 \pm 12$

Obviously, the increase has the highest effect on the lowest reported uncertainties. One gets new characteristics:

$x_w = 101.6$
$s_{ext} = 6.4$
$s_{int} \rightarrow s_w = 6.4$
$s_u = 8.1$

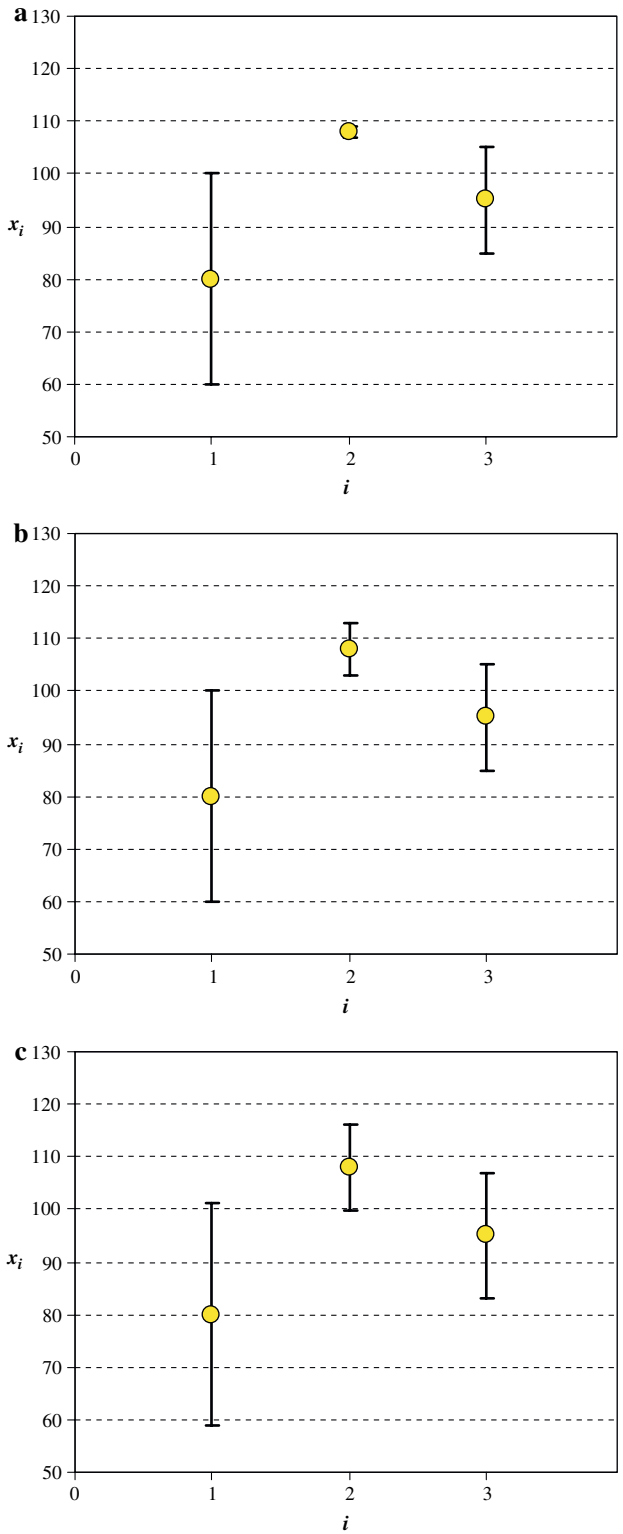


Fig. 1 **a** Hypothetical measurement results x_i provided by the n expert laboratories ($i = 1,...,n$). *Error bars* refer to standard uncertainty. **b** Data after correction in stage 1. **c** Data after adjustment of uncertainties in stage 2

The ILC coordinator decides not to increase the uncertainties even more to include a possible systematic error, as he suspects that the main problem was in the incompleteness of the (random components of the) uncertainty budget.

Stage 3

The (partially) weighted mean and uncertainty are calculated for different α-values (see Fig. 2a).

$\alpha = 0$: $x_{ref} \pm u(x_{ref}) = 94 \pm 8$
$\alpha = 1$: $x_{ref} \pm u(x_{ref}) = 98 \pm 7$
$\alpha = 2$: $x_{ref} \pm u(x_{ref}) = 102 \pm 6$

The uncertainty on x_{ref} does not vary much as a function of α because a significant value of a had to be added to the experimental uncertainties in stage 2, in order to reach the condition $R_B = 1$. The ILC coordinator has moderate trust in the relative uncertainty and takes the partial weighting result for $\alpha = 1$.

What if...?

- What if one does not correct for the uncertainty in stage 1?

Then $R_B = 1.34$, one increases the uncertainties by $a = 7.5$ and the final result, $x_{ref} \pm u(x_{ref}) = 99 \pm 7$ (for $\alpha = 1$), is still comparable to the previous result, $x_{ref} \pm u(x_{ref}) = 98 \pm 7$ (for $\alpha = 1$).

- What if one does not add a constant uncertainty a in stage 2?

There is a significant difference between weighted and unweighted mean (Fig. 2b).

$\alpha = 0$: $x_{ref} \pm u(x_{ref}) = 94 \pm 8$
$\alpha = 1$: $x_{ref} \pm u(x_{ref}) = 100 \pm 6$
$\alpha = 2$: $x_{ref} \pm u(x_{ref}) = 104 \pm 4$

One finds that the uncertainties are lower, in particular when weighting is applied. The latter is because then, the result relies more on the most 'accurate' data. The risk for underestimation of the uncertainty becomes quite high. Clearly, the 'weighted' uncertainties for $a = 6.5$ are more conservative, hence less likely to be underestimated.

- What if the methods have a common (systematic) uncertainty component?

The uncertainty budget of the measurements contains independent (random) components as well as common (systematic) uncertainty components. When checking for consistency, one should compare the data scatter with the random components only. If necessary, this part of the uncertainty should be increased to reach internal consistency. The adapted random variance propagates with a reduction factor equal to the number of measurements. The common uncertainty components, on the other hand, are added entirely and independently to the uncertainty of the reference value (see e.g. ref. [13]).

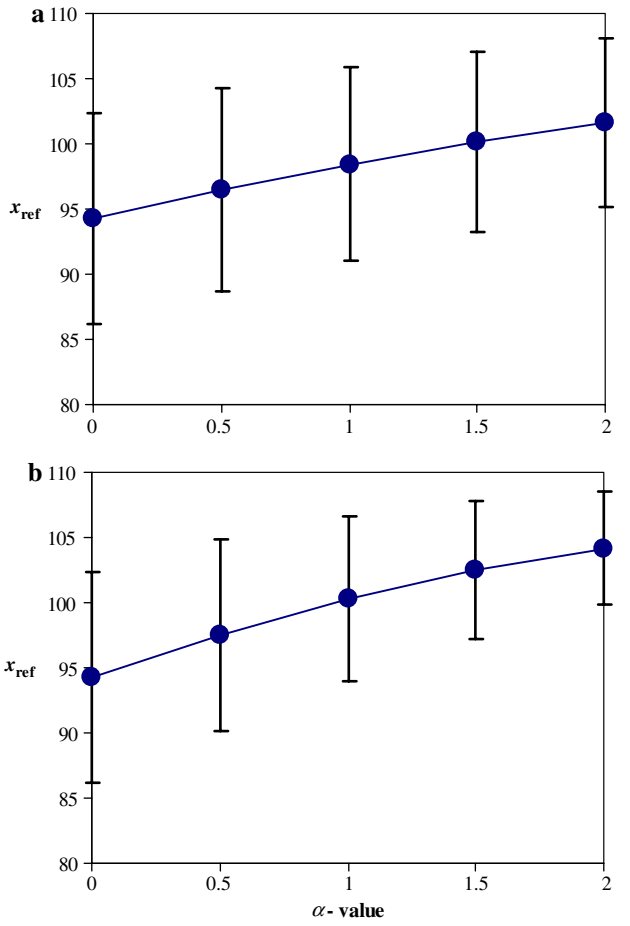


Fig. 2 **a** Calculated reference values, x_{ref}, and uncertainties for the data in Fig. 1, applying different α-values in stage 3; $\alpha = 0$ corresponds to the unweighted mean and $\alpha = 2$ to a (fully) weighted mean. **b** Same as above, assuming that in stage 2, the uncertainty has not been artificially increased ($a = 0$)

- What if we combine several results from different methods?

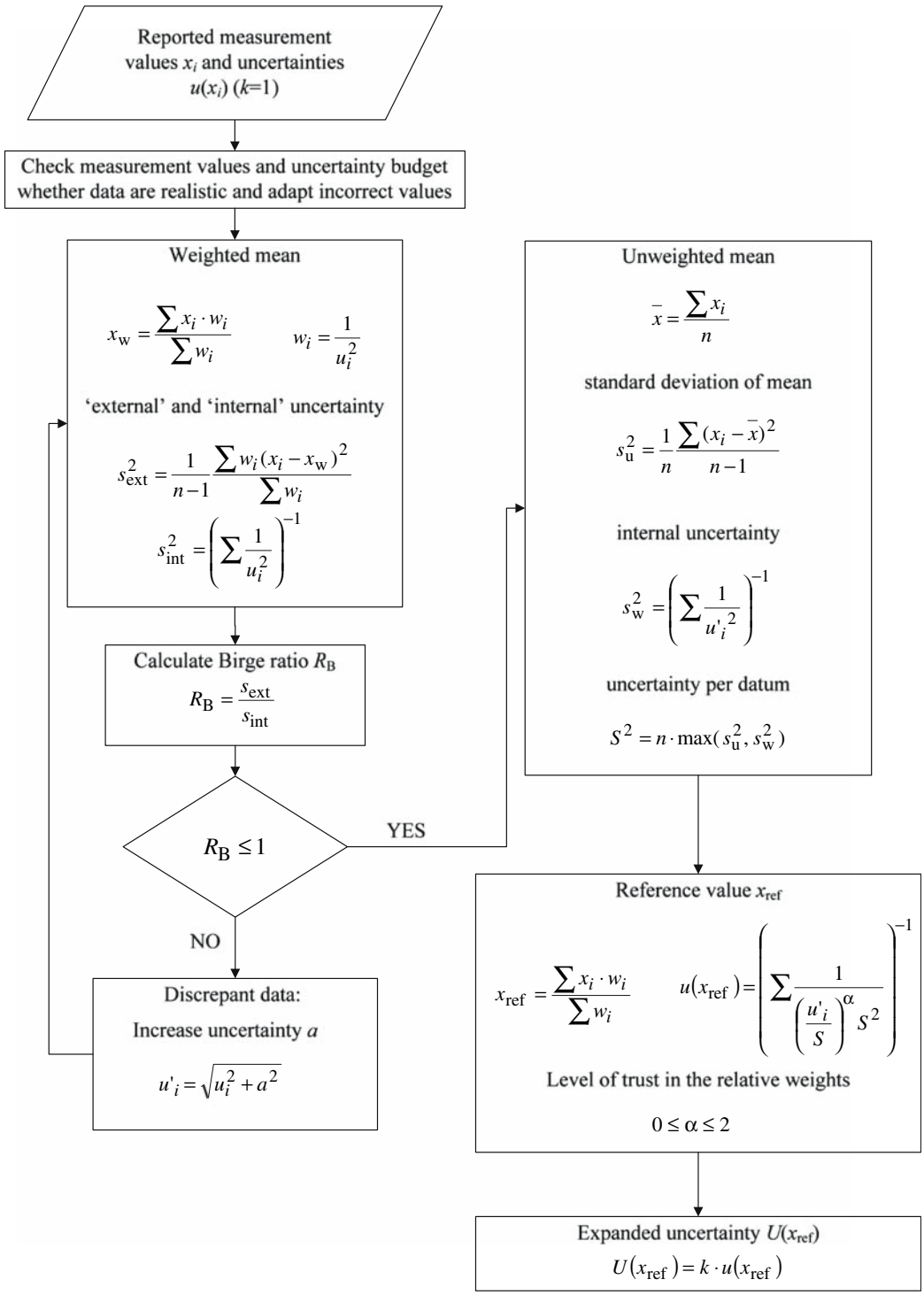
The data can be treated in groups, one group combining the results of similar methods (see also ref. [14]). A single value and uncertainty can be calculated for each group and then be treated as independent values. By doing proper uncertainty propagation for the mean within and among the groups, one will also take into account the reduction in uncertainty that was realised by having redundancy of measurements.

Conclusions

The proposed procedure yields in a harmonised way, a statistically acceptable reference value and uncertainty from a few experimental data. Yet, it leaves room for interpretation and scrutiny of the data by expert metrologists, as well as freedom in the choice of relative weighting of the data.

Appendix

Flowchart



References

1. ISO/IEC (1997) ISO guide 43-1, Proficiency testing by interlaboratory comparisons. Part I: Development and operation of proficiency testing schemes, Geneva, Switzerland
2. ISO (2005) International standard ISO 13528. Statistical methods for use in proficiency testing by interlaboratory comparisons, Geneva, Switzerland
3. Thompson M, Ellison S, Wood R (2006) Pure Appl Chem 78:145–196
4. Kacker R, Datla R, Parr A (2002) Metrologia 39:279–293
5. Pommé S (2006) Appl Radiat Isot 64:1158–1162
6. Pommé S (2006) Applied modeling and computations in nuclear science. Semkow TM, Pommé S, Jerome SM, Strom DJ (eds) ACS symposium series 945, American Chemical Society, pp 282–292, ISBN 0-8412-3982-7
7. Cox MG (2007) Metrologia 44:187–200
8. Lira I (2007) Metrologia 44:415–421
9. Cox MG, Forbes AB, Flowers JL, Harris PM (2004) Advanced mathematical and computational tools in metrology VI. Ciarlini P, Cox MG, Pavese F, Rossi GB (eds) Series on advances in mathematics for applied sciences, vol 66, World Scientific, pp 37–51
10. Cox MG (2002) Metrologia 39:589–595
11. Pennecchi F, Callegaro L (2006) Metrologia 43:213–219
12. Willink R (2007) Metrologia 44:105–110
13. Pauwels J, Lamberty A, Schimmel H (1998) Accred Qual Assur 3:180–184
14. Levenson MS, Banks DL, Eberhardt KR, Gill LM, Guthrie WF, Liu HK, Vangel MG, Yen JH, Zhang NF (2000) J Res Natl Inst Stand Technol 105:571–579