Determination of a reference value and its uncertainty through a power-moderated mean

S Pommé1 and J Keightley2

1 European Commission, Joint Research Centre, Institute for Reference Materials and Measurements, Retieseweg 111, B-2440 Geel, Belgium
2 National Physical Laboratory, Hampton Road, Teddington, Middlesex TW11 0LW, UK

E-mail: stefaan.pomme@ec.europa.eu

Received 18 March 2015, revised 1 April 2015
Accepted for publication 2 April 2015
Published 22 May 2015

Abstract
A method is presented for calculating a key comparison reference value (KCRV) and its associated standard uncertainty. The method allows for technical scrutiny of data, correction or exclusion of extreme data, but above all uses a power-moderated mean that can calculate an efficient and robust mean from any data set. For mutually consistent data, the method approaches a weighted mean, the weights being the reciprocals of the variances (squared standard uncertainties) associated with the measured values. For data sets suspected of inconsistency, the weighting is moderated by increasing the laboratory variances by a common amount and/or decreasing the power of the weighting factors. By using computer simulations, it is shown that the PMM is a good compromise between efficiency and robustness, while also providing a realistic uncertainty. The method is of particular interest to data evaluators and organizers of proficiency tests.

Keywords: KCRV, mean, power-moderated mean, uncertainty, outlier, degree of equivalence, data evaluation

Online supplementary data available from stacks.iop.org/MET/52/03S200/mmedia

(Some figures may appear in colour only in the online journal)

1. Introduction

Data evaluators have the task to derive the best possible estimate of a measurand from a set of N measurement data \( x_i \) and associated standard uncertainties \( u_i \). Often they are faced with discrepant data sets, for which the stated uncertainties do not cover the observed dispersion of data. In this community, there is a need for an estimator that provides an efficient mean and realistic uncertainty, which are robust against data that are extreme in value and/or uncertainty. In a trade-off between efficiency and robustness of the mean, the quality of the uncertainty evaluations by the metrologists plays a crucial role: efficiency can be gained through statistical weighting of data provided that the stated \( u_i \) are realistic, whereas robustness demands for countermeasures where they are not.

This problem is shared by organizers of proficiency tests, and ultimately by Consultative Committees (CCs) of the International Committee for Weights and Measures (CIPM). In the frame of Key Comparisons (KCs), a reference value (KCRV) is evaluated from the data sets \( (x_i, u_i) \) obtained from \( N \) participating laboratories. For decades, the uncertainties \( u_i \) have generally been disregarded for the calculation of a KCRV, the KCRV being calculated as an arithmetic mean. In May 2013, the Consultative Committee for Ionizing Radiation, Section II: Measurement of radionuclides—or CCRI(II)—has adopted a new method to calculate a KCRV that accounts for the \( u_i \). The method applies to data which are mutually independent and normally distributed around the same value.
It is based on a few fundamental principles:

- The estimator should be efficient, providing an accurate KCRV on the basis of the available data set \((x_i, u_i)\) and technical scrutiny.
- The estimator should give a realistic standard uncertainty on the KCRV.
- The \(N\) data are treated on an equal footing, although relative weighting may vary as a function of stated uncertainty. The method shall optimize the use of information contained in the data.
- Before evaluation, all data is scrutinized in an initial data screening by (representatives of) the CC, which may choose to exclude from the evaluation of the KCRV or correct data on technical grounds.
- Extreme data can be excluded from the evaluation of the KCRV on statistical grounds as part of the method. The CC is always the final arbiter regarding excluding any data from the evaluation of the KCRV.
- The estimator should be robust against extreme data, in case such data have not been excluded from the data set. It should also adequately cope with discrepant data sets.
- The method is preferably uncomplicated and conveniently reproducible.

The search for appropriate methods to analyse intercomparison data has in recent years been a major topic in metrology-related journals (see e.g. references in [1, 2]). At the heart of the problem lies the fact that the stated uncertainty values \(u_i\) are generally imperfect estimates of the combined effect of all sources of variability and are therefore susceptible to error. A study of the CCRI(II) data in the Key Comparison Data Base (KCDB) revealed that (i) the data sets often show signs of discrepancy and that (ii) there is a positive correlation between stated uncertainties \(u_i\) and mean absolute deviations from the reference value \(x_i = \text{KCRV}\), but (iii) the relationship between both is not linear [3]. It was concluded that, for such data sets, a value proportional to \(u_i^{-1}\) would be more appropriate as relative weighting factor for the mean than the reciprocal of the variance \(u_i^{-2}\) [3].

The estimator chosen by the CCRI(II) is the power-modified mean (PMM) [4], based on a concept by Pommé and Spasova [5]. It can be regarded as an upgrade of the well-established Mandel–Paule (M–P) mean [6, 7]. Its results are generally intermediate between arithmetic and weighted mean. The method combines several notions also expressed by others: (i) to scrutinize and possibly correct data prior to calculating a mean [8], (ii) to test data for consistency on the basis of a Birge ratio or \(\chi^2\) [6–10], (iii) to exclude extreme data on a statistical basis [11–13], (iv) to adjust uncertainties to reach consistency [6, 7, 14, 15], (v) to reduce the power to which uncertainties are used as weighting factors [3, 11], (vi) to test the adequacy of the estimator by simulation of discrepant data sets [16], (vii) to tailor the estimator to the quality of the data [17].

The procedure is discussed in this paper. Section 2 explains step by step how the KCRV and its uncertainty are calculated from the PMM. Section 3 considers the identification and treatment of data regarded as statistically extreme. Section 4 describes the determination of degrees of equivalence. Section 5 summarizes the method and section 6 shows some examples. Section 7 gives conclusions.

Annex A gives a resume of the rationale behind the choice of estimator on the basis of conclusions drawn from simulations. Annex B contains an overview of relevant formulae for the arithmetic mean, the classical weighted mean, the M–P mean, the DerSimonian–Laird (D–L) mean and the PMM. In annex C, formulae are derived for degree of equivalence and outlier identification. An excel spreadsheet with implementation of the power-modulated mean is available online as supplementary data. (stacks.iop.org/MET/52/03S200/mmmedia)

2. The power-modernated weighted mean

The estimator combines aspects of the arithmetic mean, the weighted mean and the M–P mean. The logical steps leading to this procedure can be read in annex B.

In this section, the mathematical steps are shown in order of execution:

(a) Calculate the M–P mean

\[
    x_{\text{mp}} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} \frac{1}{u_i^2 + s^2}},
\]

using \(s^2 = 0\) as initial value, which conforms to the weighted mean.

(b) Calculate the modified reduced observed \(\chi^2\) value

\[
    \tilde{\chi}^2_{\text{obs}} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - x_{\text{mp}})^2}{u_i^2 + s^2},
\]

(c) If \(\tilde{\chi}^2_{\text{obs}} > 1\), increase the variance \(s^2\) and repeat steps 1–2 until \(\tilde{\chi}^2_{\text{obs}} = 1\) is obtained.

(d) Assess the reliability of the uncertainties provided. Choose a value for the power \(\alpha\) of the uncertainties in the weighting factors (see annex B.4):

| Power | Reliability of uncertainties |
|-------|-----------------------------|
| \(\alpha = 0\) | the uncertainties \(u_i\) are uninformative, disproportional to errors on \(x_i\) (arithmetic mean) |
| \(\alpha = 0\) | the uncertainties \(u_i\) are weakly proportional to errors on \(x_i\), but the variation of \(u_i\) due to errors in the uncertainty evaluation is at least twice as large as due to differences in metrological accuracy (arithmetic mean) |
| \(\alpha = 2–3/N\) | informative but imperfect uncertainties, uncertainties with a tendency of being underestimated (intermediately weighted mean) |
| \(\alpha = 2\) | realistic uncertainties with a modest error; no specific trend of underestimation, large data sets (M–P mean) |
| \(\alpha = 2\) | accurately known uncertainties, consistent data (weighted mean) |
(e) Calculate a characteristic uncertainty per datum, based on the variance associated with the arithmetic mean, \( x \), or the M–P mean, \( x_{\text{mp}} \), whichever is larger.\(^3\)

\[
S = \sqrt{N \cdot \max(\sigma^2(x), \sigma^2(x_{\text{mp}}))},
\]

in which

\[
u^2(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N(N-1)}, \quad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i,
\]

and

\[
u^2(x_{\text{mp}}) = \left( \sum_{i=1}^{N} \frac{1}{u_i^2 + s^2} \right)^{-1}.
\]

(f) Calculate the reference value and uncertainty from a power-moderated weighted mean

\[
x_{\text{ref}} = \sum_{i=1}^{N} w_i x_i + \frac{1}{\nu^2(x_{\text{ref}})} = \sum_{i=1}^{N} \left[ \left( \nu(x)^2 + s^2 \right)^{\alpha} \nu^2 - \alpha \right]^{-1},
\]

in which the normalized weighting factor is

\[
w_i = \nu^2(x_{\text{ref}}) \left( \nu(x)^2 + s^2 \right)^{\alpha} \nu^2 - \alpha \]

3. Treatment of extreme data

Statistical tools may be used to indicate data that are extreme. An extreme datum is such that the magnitude of the difference \( e_i \) between a measured value \( x_i \) and a candidate KCRV \( x_{\text{ref}} \) exceeds a multiple of the standard uncertainty \( u(e_i) \) associated with \( e_i \):

\[
|e_i| > ku(e_i), \quad e_i = x_i - x_{\text{ref}},
\]

where \( k \) is a coverage factor, typically between two and four, corresponding to a specified level of confidence.

For a weighted mean, the equations for the variance of the difference can be expressed through the normalized weighting factors (see annex C.1). Applying the same equations (8) and (9) to the PMM provides an elegant way to use the modified uncertainties:

\[
u^2(e_i) = \nu^2(x_{\text{ref}}) \left( \frac{1}{w_i} - 1 \right) \quad (x_i \text{ incl. in mean}), \quad (8)
\]

\[
u^2(e_i) = \nu^2(x_{\text{ref}}) \left( \frac{1}{w_i} + 1 \right) \quad (x_i \text{ excl. from mean}), \quad (9)
\]

The approach of using the modified uncertainties limits the number of values for which the inequality in expression (7) holds. Preferably, the identification and rejection of extreme data is kept to a minimum, so that the mean is based on a large subset of the available data. A default coverage factor of \( k = 2.5 \) is recommended, assuming that the data with augmented uncertainties are approximately normally distributed around the mean.

\(^3\) Both variances are equal when \( \tilde{\chi}_{\text{obs}}^2 = 1 \).

4. Degree of equivalence

The degree of equivalence, DoE, for the \( i \)th laboratory has two components \((d_i, U(d_i))\), where, assuming normality,

\[
d_i = x_i - x_{\text{ref}}, \quad U(d_i) = 2u(d_i).
\]

\( u(d_i) \) is the standard uncertainty associated with the value component \( d_i \), and \( U(d_i) \), the uncertainty component, is the expanded uncertainty at the 95% level of confidence.

Given a KCRV \( x_{\text{ref}} \) and the associated standard uncertainty \( u(x_{\text{ref}}) \) obtained from expressions (5) and (6), the corresponding DoEs are determined from the generally valid expression for any kind of weighted mean\(^4\) (see annex C or [18]):

\[
u^2(d_i) = (1 - 2w_i)u_i^2 + \nu^2(x_{\text{ref}}).
\]

The DoEs for participants whose data were excluded from the calculation of the KCRV are given by essentially the same expression, applying \( w_i = 0 \):

\[
u^2(d_i) = u_i^2 + \nu^2(x_{\text{ref}}).
\]

The variance \( u_i^2 \) associated with \( x_i \) is not augmented by \( s^2 \) for the calculation of the DoE, since it is the measurement capability of laboratory \( i \), including proper uncertainty statement, that is being assessed.

5. Summary of the method

(a) Carry out a careful examination of the participants’ data. If necessary, correct or exclude erroneous data on technical grounds.

(b) Form the weighted mean and the associated standard uncertainty of the remaining data, \((x_i, u_i), i = 1, \ldots, N\), using equation (1) with \( s^2 = 0 \).

(c) Test for consistency of the data with the weighted mean by calculating \( \tilde{\chi}_{\text{obs}}^2 \) using equation (2), regarding the data as consistent if \( \tilde{\chi}_{\text{obs}}^2 \leq 1 \).

(d) If \( \tilde{\chi}_{\text{obs}}^2 > 1 \), calculate the M–P mean of the remaining data. That is, the variance \( s^2 \) in the weighted mean (equation (1)) is chosen such that \( \tilde{\chi}_{\text{obs}}^2 \) (equation (2)) is unity.

(e) Choose a value for the power \( \alpha \) based on the reliability of the uncertainties and the sample size.

\(^4\) In the particular case of the classical weighted mean, the expression for \( u^2(d_i) \) is identical to \( u^2(x_{\text{ref}}) \), because \( w_i = \nu^2(x_{\text{ref}})u_i^2 \). The more general expression has to be used for the M–P mean and PMM (\( s^2 > 0 \) or \( \alpha < 2 \)).
(f) Calculate the PMM and its standard uncertainty from equations (3)–(6).

(g) Use the statistical criterion in equations (7)–(9) to identify any further extreme values, applying the normalized weighting factors in equation (6). Should the CC exclude such data from the calculation of the KCRV, repeat steps 2–6 on the remaining data set.

(h) Take the PMM as the KCRV and its associated standard uncertainty as the standard uncertainty associated with the KCRV.

(i) Form the DoEs for all participating laboratories using equations (10)–(12).

6. Examples

Figure 1 shows the PMM and its uncertainty for four data sets. Data set A is consistent and the PMM approaches the weighted mean, even though using a reduced power of only 1.25 ($N = 4$). The uncertainty on the arithmetic mean would be unrealistically low if based on the data variance only. Data set B contains a potential outlier, inclusion or exclusion of which are both acceptable within the criteria set in equations (7)–(9) ($k = 2.5$), thus leading to two possible solutions. If the extreme datum is excluded from the initial calculation of the PMM (cf. step 1 of the procedure), the PMM has a small uncertainty, and using equation (9) the extreme datum exceeds the criterion in equation (7). If all data are included in the PMM, the uncertainty $u(x_{ef})$ is 2.6 times higher, using equation (8) the extreme datum passes the criterion, and the PMM shifts by almost one standard uncertainty. Data set C is dominated by one datum with tiny uncertainty, in particular when applying a weighted or M–P mean. The PMM moderates the influence of this single datum, leading to an intermediate value with an uncertainty higher than that of the dominant datum. Data set D is not consistent in the sense of the criterion $\chi^2_{2\alpha} \leq 1$. The uncertainties are increased and the PMM does not deviate much from the M–P mean.

7. Conclusions

The power-moderated mean is an efficient and robust estimator of the reference value of a data set and its uncertainty. The method applies when the measured values are mutually independent and normally distributed around the same value. The PMM takes the evaluated uncertainties of all data into account in the weighting, but the relative weighting factors are adjusted to the level of consistency in the data set. Replacing the arithmetic mean by the PMM in CCRI(II) key comparisons has led to a more consistent set of KCRVs. Simulations confirm that the PMM is more robust than a DerSimonian–Laird or Mandel–Paule mean.

The method provides protection against erroneous and extreme data through the possibility for correction or exclusion of data. Further discrepancy of the data, most typically caused by understatement of the uncertainty, is generally well taken into account by the estimator. This is established by temporarily increasing the uncertainties and reducing the power of uncertainties in the weighting factors. It is performed purely for the calculation of the KCRV, as the laboratory data remain unaltered when obtaining degrees of equivalence.

For consistent data with realistic uncertainties, the KCRV approaches the classical weighted mean. For highly discrepant data with uninformative uncertainties, the KCRV approaches the arithmetic mean. There is a smooth transition from the weighted mean to the arithmetic mean as the degree of data inconsistency increases. For CCRI(II) intercomparison results, typically slightly discrepant data with informative but imperfectly evaluated uncertainties, the KCRV is intermediate between the Mandel–Paule mean and the arithmetic mean.

Simulations suggest that the PMM is approximately normally distributed around the true value and its uncertainty is a good estimate for the distribution width. The method remains vulnerable to seemingly consistent data sets with a common bias or with systematically understated (or overstated) uncertainties.

Acknowledgments

The authors are indebted to Professor Dr Maurice Cox (NPL), Dr Peter Harris (NPL) and Dr Carine Michotte (BIPM) for their active involvement in discussions on this topic within the KCWG of the CCRI(II). Special thanks are due to Dr Penelope Allisy-Roberts (BIPM) for years of outstanding work as the Executive Secretary of CCRI, hosting the meetings and embracing innovative ideas.
Annex A. Rationale behind the choice of estimator

A.1. Conclusions from literature

Measurement results show deviations from the ‘true’ value of the measurand. The same holds for the reported uncertainty which, in general, is only a rough estimate of the combined effect of all sources of variability. For example, the international comparison CCRI(II)-S7 [19] on the analysis of uncertainty budgets for 4πβγ coincidence counting showed differences up to a factor of 3 on the evaluated combined uncertainty of a common data set and even larger diversity in the composition of the uncertainty budget. Since the weighting factors \( w_i = u_i^{-2} \) in this exercise would differ by a factor of 9, the weighted mean would be (unjustly) drawn towards the result with the lowest (and possibly the least realistic) uncertainty. Intercomparison results presented in PomPlots [3, 20] show recurring symptoms of underestimation of uncertainties, often imbedded in the culture of laboratories, and in many cases the relative differences among the stated uncertainties \( u_i \) are larger than the differences in accuracy of the measurement result \( x_i \). Surprisingly, statistical paradigms seem to focus more on unexplained biases \( e_i \) in \( x_i \), instead of the large errors on \( u_i \) and their impact on data (in)consistency.

Duewer [16] performed extensive computer simulations to evaluate the performance of a suite of estimators of the mean applied to data sets contaminated with value and uncertainty outliers. Important criteria were efficiency, a measure of the expected variance of the mean for repeated samplings of the population, and robustness of the mean and its uncertainty against extreme data, a measure of insensitivity to violations of the assumption that the data are distributed with a mean and standard deviation equal to the true value and the stated uncertainty, respectively.

Some conclusions that can be derived from the work of Duewer [16] are

(a) If none of the \( u_i \) is informative, the arithmetic mean is the most efficient estimator.
(b) If the uncertainties \( u_i \) are informative, an estimator that uses them can be employed to improve the efficiency. Some approaches using the \( u_i \) are better than others.
(c) The classical weighted mean, using the reciprocal of the variances as weighting factors, is the most efficient estimator for normally distributed data, only in absence of unrepresentative data. Even modest contamination by such data, in particular those having extreme values and/or understated uncertainties, results in an uncertainty estimate that is too low.
(d) The Mandel–Paule (M–P) mean provides a good combination of efficiency and robustness for discrepant data that are approximately symmetrically disposed with respect to the KCRV. It degrades little with increasing contamination, is only slightly dependent on the level of information carried by the \( u_i \), and is superior to the classical weighted mean when the \( u_i \) are not very informative.

An important conclusion of Duewer’s work was that the M–P mean (referred to as ‘wMean’) provides ‘the best combination of efficiency and location- and M[measurement] U[uncertainty]-robustness for symmetric contamination’ [16]. The PMM inherits these good characteristics, and two features were added to improve its robustness: a moderation of the power in the weighting factors and an outlier rejection criterion.

In this work, additional simulations were performed on discrepant data sets for which the variation of the data \( x_i \) exceeds expectation from the stated uncertainties \( u_i \). The PMM is compared with three reference methods also evaluated by Duewer: the arithmetic mean, the weighted mean and the M–P mean. A comparison is also made with the DerSimonian–Laird (D–L) procedure [21, 22], a widely used estimator for meta-analysis in clinical trials. A selection of the examined data is discussed in section A.2. The PMM has also been applied to the inter-comparison data sets of CCRI(II) in the KDCB. Examples and conclusions are discussed in section A.3. The applied equations for the estimators are summarized in annex B.

A.2. Computer simulations

Data were generated from normal distributions with the same mean and different widths:

\[ x_i \sim \mathcal{N}(\mu, \sigma), \quad (A.1) \]

in which for each \( x_i \) the ‘true’ standard deviation \( \sigma \) is varied around a mean value \( \mu \) by means of a randomly generated multiplication factor \( f_m \) representing differences in metrological precision. An additional factor \( f_e \) was generated to simulate the hypothetical errors made in uncertainty evaluation, including a ‘shift’ factor \( F_e \) to impose a tendency for underestimation (or overestimation) of uncertainties. Data sets comprising of 2–20 \((x_i, u_i)\) pairs were generated in which

\[ u_i = f_m f_e u, \quad (A.2) \]

and \( f_m = F_m^{2(X-0.5)}, f_e = F_e^{2(Y-0.5)} F X, Y \sim U([0,1]) \). Using an amplitude of e.g. \( F_m = 3 \), the numbers generated for \( f_m \) vary between \( 1/3 \) and 3. Applying e.g. \( F_e = 0.7 \) and \( F_e = 3 \) generates numbers for \( f_e \) from 1/4.3 to 2.1. In all simulations \( \sigma = N^{1/2} \) was chosen.

The efficiency of the method is quantified through the standard deviation of the mean relative to the arithmetic mean (equation (B.4)), \( \sigma(x_{\text{ref}})/\sigma(\text{AM}) \), both values averaged over many samplings. The robustness of the uncertainty of the mean \( u(x_{\text{ref}}) \) against erroneous input uncertainty values is evaluated from the ratio between the standard deviation of the mean and the estimated uncertainty, \( \sigma(x_{\text{ref}})/u(x_{\text{ref}}) \), both values averaged over many samplings.

Simulations were performed for different ratios between \( F_m \) and \( F_e \), i.e. the variation of \( u_i \) values due to ‘true’ metrological differences in the uncertainties and ‘errors’ in the uncertainty evaluation: \( F_m/F_e = 2/4, 3/3, 4/2 \) and \( 4/1 \), while applying \( F_e = 1 \) (symmetric discrepancies in uncertainties \( u_i \)) or \( F_e = 0.7 \) (tendency for underestimation of uncertainties).

Figure 2 shows the ‘efficiency’ and ‘robustness’ for a scenario in which \( F_m \) and \( F_e \) are of comparable size \((F_m/F_e = 3/3)\) and the uncertainties are biased towards low values \((F_e = 0.7)\).
The data sets are discrepant and the reduced chi-squared has values typically around $\chi^2_{\text{obs}} \approx 3-4$. As expected, the efficiency is the lowest for the AM, significantly better for the PMM ($\alpha = 2-3/3$), M–P and D–L, and the highest for the WM. On the other hand, the WM is not robust since its uncertainty $u(\text{WM})$ is too small compared to the variation $\sigma(\text{WM})$. The AM uncertainty is more realistic and also the M–P uncertainty performs well for large data sets. The D–L uncertainty is systematically low, and not only for small data sets as anticipated from published simulation results [23]. Compared to D–L and M–P, the PMM uncertainty is generally larger, in particular for small data sets, and leads to fewer cases of underestimation of uncertainty. These results confirm that for discrepant data with informative uncertainties, the PMM yields the best compromise between efficiency and robustness.

In this configuration ($F_{\text{ref}}/F_{\text{e}} = 3/3$, $F_{\text{e}} = 0.7$, $N = 4$), the PMM is roughly normally distributed with width $u(x_{\text{ref}})$ around the true value $\mu$, since the probability of finding $\mu$ within $x_{\text{ref}} \pm ku(x_{\text{ref}})$ is 67%, 26%, 5% and 1.2% for $k = 1, 2, 3, 4$ respectively. The fraction of cases in which $|\mu - x_{\text{ref}}| > 4u(x_{\text{ref}})$ is smaller for the PMM (0.4%) than for the AM (1.0%) (using equation (B.4) for the uncertainty), the D–L mean (2.5%), the M–P mean (2.6%) and the WM (12%). For large data sets ($N = 20$) there are generally fewer outliers: AM (0.02%), PMM (0.16%), M–P mean (0.24%), D–L (0.31%), except for the WM (17%).

Figure 3 presents data sets in which the uncertainties are underestimated and their variation due to errors is twice the variation due to metrological differences, i.e. $F_{\text{e}}/F_{\text{ref}} = 2/4$ and $F_{\text{e}} = 0.7$. Here the WM is the least efficient estimator due to its sensitivity to a small subset of dominant data points. Also the weighting in the D–L, M–P and PMM ($\alpha = 2-3/3$) estimators offers no significant efficiency gain compared to the AM, which remains the best choice for data with uninformative uncertainties. Even safer alternatives are the PMM with $\alpha = 0$ or AM using uncertainty equation (B.5) instead of equation (B.4).

For the data sets in figure 4 with unbiased and nearly optimal uncertainties, i.e. $F_{\text{e}}/F_{\text{ref}} = 4/2$ and $F_{\text{e}} = 1$, the uncertainty weighting has a very advantageous effect on the efficiency. In spite of the high quality of the input data, there is still a trend of underestimation of the uncertainty of the WM, D–L and M–P mean, whereas the PMM ($\alpha = 2-3/3$) on average slightly overestimates the uncertainty. The latter is intentional and not harmful, since the PMM uses the maximum between the uncertainties of the AM and the M–P mean in the parameter $S^2$.

Figure 2. Measures for efficiency and robustness of the arithmetic, weighted, Mandel–Paule, Der Simonian-Laird and power-moderated mean for discrepant data sets ($F_{\text{ref}} = 3, F_{\text{e}} = 3, F_{\text{e}} = 0.7$) as a function of the number of data. (Top) Standard deviation of the mean around the true value used in the simulation. (Bottom) The ratio of the standard deviation of the mean to the calculated uncertainty.

Figure 3. The same as in figure 2 for data sets with ‘uninformative’ uncertainties, i.e. quite randomly estimated uncertainties with a tendency towards low values ($F_{\text{ref}} = 2, F_{\text{e}} = 4, F_{\text{e}} = 0.7$).
in equation (3) (or (B.11)). A similar increase of the average AM uncertainty would be obtained if equation (B.5) were used instead of equation (B.4), which is nevertheless recommendable to avoid individual cases of severe underestimation of \( u(AM) \) for data sets with a low variance caused by chance (see figure 1(A) for an example).

Figure 5 shows simulation results for ideal data sets with correct uncertainties \( (F_m = 4, F_e = 2, F_s = 1) \). All estimators perform adequately and their efficiency is inversely correlated with the average uncertainty.

Binned frequency distributions of \( k = |\mu - x_{\text{ref}}|/u(x_{\text{ref}}) \) for the PMM, the D–L mean and the WM are shown in figures 6–8, respectively, for data sets of size \( N=5 \) at the four simulated levels of accuracy in the uncertainty \( u_i \). For data sets with imperfect but informative uncertainties, the PMM is nearly normally distributed around the true value, whereas the WM and D–L mean have more outliers.

The main conclusions drawn from the simulations are

(a) For very discrepant data sets, the PMM, M–P and D–L estimates approach the arithmetic mean. For consistent data sets, they approach the classical weighted mean.
(b) Using informative uncertainties in the weighting factor increases the efficiency of the WM, D–L, M–P, PMM estimators. Uninformative uncertainties have the opposite effect.
(c) If the variation of uncertainties \( u_i \) due to errors in the uncertainty assessment is twice or more the variation due to metrological origin \( (F_e > 2F_m) \), the arithmetic mean is the most efficient, or the PMM with \( \alpha = 0 \).
(d) Applied to data sets contaminated with underestimated uncertainties, the WM, D–L and M–P estimators to various degrees result in an uncertainty estimate that is too low.
(e) The power-moderated mean yields more reliable uncertainties for discrepant data sets in which uncertainties are likely to be underestimated. It uses moderate weighting for small data sets, thus being less influenced by unidentified outliers with underestimated uncertainties.

A.3. Application to data in the KCDB

The PMM was calculated for the BIPM.RI(II)-K1 key comparison data in the KCDB, and compared with the arithmetic mean, the weighted median [24] and the weighted mean of the largest consistent subset of each data set [13]. Up to the year 2013, the KCRVs were calculated from the arithmetic mean and outliers were ‘if necessary, excluded from the KCRV using the normalized error test with a test value of four’.

The weighted median and weighted mean were overly influenced by the data with the lowest uncertainties and
led to underestimates of the KCRV uncertainty. The WM over the largest consistent subset was not representative of all available results and offered no solution for small discrepant data sets. The PMM solved a few flaws of the AM: low efficiency, underestimation of the uncertainty for consistent data sets, and overweighting of data with large uncertainty. In fact, there was some ambiguity about the method by which the uncertainty $u(AM)$ was calculated: the squared sum of the stated uncertainties was used (equation (B.2)) for the calculation of the DoEs [18], whereas the uncertainty of the KCRV was calculated from the standard deviation of the data set (equation (B.3)).

Figure 9 shows key comparison data for the activity of a $^{99m}$Tc solution. Due to the good agreement of the results, the uncertainty of the AM (derived from equation (B.3)) was underestimated as being even lower than the uncertainty of a WM. The combined uncertainties were correctly taken into account by the PMM. In figure 10 key comparison data are shown for the activity of a $^{65}$Zn solution. Both the AM and PMM (equation (7), $k = 2.5$) identified $i = 11$ as a potential outlier. The AM was biased towards the low value of $i = 10$, due to the equal weight assigned to all data. The PMM automatically produced a more realistic KCRV by assigning a low weight to $i = 10$ due to its high uncertainty $u_i$.

All in all, the KCRVs derived from the PMM showed better consistency with the predicted values from the calculated efficiency of the SIR (Système International de Référence) [25].

Figure 6. Binned frequency distribution of $k = \left| \mu - x_{\text{ref}} \right| / u(x_{\text{ref}})$ for the power-moderated mean applied to the simulated data in figures 2–5, compared to the normal distribution.

Figure 7. The same as figure 6 for the D–L estimator.
Conclusions from the practical use of these estimators are:

(a) The outlier identification mechanism efficiently protects the PMM from the most extreme data, whilst being inclusive for the majority of data in discrepant as well as in consistent sets.

(b) The AM and PMM are more representative of a discrepant data set than a WM on the largest consistent subset.

(c) The weighted median and WM are highly influenced by data with a small uncertainty.

(d) The AM has issues with low efficiency, sensitivity to extreme data with large uncertainty, underestimation of uncertainty (depending on which formula is used).

Annex B. Formulae for KCRV estimators and associated standard uncertainties

B.1. Arithmetic mean

The arithmetic mean is calculated from

\[ \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \]  

and its uncertainty, applying the propagation rule, is

\[ u(\bar{x}) = \frac{1}{N} \sqrt{\sum_{i=1}^{N} u_i^2}. \]  

As the arithmetic mean is of particular interest when the \( u_i \) are not informative, one can replace the individual variances \( u_i^2 \) by an estimate of the sample variance

\[ u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \]  

resulting in

\[ u(\bar{x}) = \frac{1}{N} \sqrt{N u^2} = \frac{u}{\sqrt{N}} = \left( \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N(N-1)} \right)^{1/2}. \]  

As the dispersion of data is the consequence of a random process, the calculated uncertainty of the mean can sometimes be unrealistically low; a particular case is small data sets showing almost no scatter. If the \( u_i \) are informative with respect to the uncertainty scale, one could take the maximum value from both approaches:

\[ u(\bar{x}) = \max \left( \frac{1}{N} \sqrt{\sum_{i=1}^{N} u_i^2}, \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N(N-1)} \right). \]  

B.2. Weighted mean

The classical weighted mean uses the reciprocal variances as weighting factors. The weighted mean \( x_{\text{ref}} \) of the data set and the associated standard uncertainty \( u(x_{\text{ref}}) \) are obtained from

\[ x_{\text{ref}} = \frac{\sum_{i=1}^{N} x_i u_i^2}{\sum_{i=1}^{N} u_i^2} \]  

or

\[ x_{\text{ref}} = u^2(x_{\text{ref}}) \sum_{i=1}^{N} \frac{x_i}{u_i^2}, \]  

and

\[ \frac{1}{u^2(x_{\text{ref}})} = \sum_{i=1}^{N} \frac{1}{u_i^2} \]  

or

\[ \frac{1}{u(x_{\text{ref}})^2} = \left( \sum_{i=1}^{N} \frac{1}{u_i^2} \right)^{-1/2}. \]

The weighted mean and its uncertainty are particularly inadequate when applied to discrepant data with understated uncertainties. One can look for indications of discrepancy by calculating the reduced observed chi-squared value.

Figure 8. The same as figure 6 for the weighted mean.
A $\chi^2_{\text{obs}}$ value (significantly) higher than unity (is) may be an indication of inconsistency.\(^5\)

\[ \chi^2_{\text{obs}} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - x_{\text{ref}})^2}{u_i^2}. \]  

(B.8)

A $\chi^2_{\text{obs}}$ value (significantly) higher than unity (is) may be an indication of inconsistency.\(^5\)

B.3. Mandel–Paule mean

The M–P mean was designed to deal with discrepant data sets, having a reduced observed chi-squared value $\chi^2_{\text{obs}}$ larger than unity. For the purpose of establishing a more robust mean, the laboratory variances $u_i^2$ are incremented by a further variance $s^2$ to give augmented variances $u_i^2 + s^2$. The value of the unexplained variance $s^2$ is chosen such that the modified reduced observed chi-squared value,

\[ \chi^2_{\text{obs}} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - x_{\text{ref}})^2}{u_i^2 + s^2}, \]  

(B.9)

equals one.\(^6\)

\(^5\)The mean (or expectation) of a variable having a chi-squared distribution with $N-1$ degrees of freedom is $N-1$. Under normality conditions, the expected value of $\chi^2_{\text{obs}}$ is unity.

\(^6\)This condition is equivalent to the ‘Birge ratio’ test as defined in [5], requiring a match between the ‘internal’ and ‘external’ variances.
The calculation of the M–P mean and its uncertainty proceeds through the same equations as for the weighted mean, replacing the stated variances \( u_i^2 \) by the augmented variances \( u^2(x_i) \)

\[
x_{\text{ref}} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} \left( \frac{1}{u_i^2 + s^2} \right)} = \sum_{i=1}^{N} \frac{1}{u_i^2 + s^2}.
\]  

(B.10)

As the M–P mean \( x_{\text{ref}} \) occurs in the equation for \( \hat{x}_{\text{obs}} \), an iterative procedure is applied to find the appropriate value of the variance \( s^2 \).

For data sets with \( \hat{x}_{\text{obs}} \) smaller than 1, the variances are not augmented, \( s^2 = 0 \), and the result is identical to the weighted mean. For an extremely inconsistent set, \( s^2 \) will be large compared with the \( u_i^2 \) and the M–P mean will approach the arithmetic mean. For intermediate cases, the influence of those laboratories that provide the smallest uncertainties will be reduced and the standard uncertainty associated with the KCRV will be larger compared with that for the weighted mean.

Though much more robust than the weighted mean, the M–P procedure tends to underestimate the uncertainty of the M–P mean for data sets with predominantly understated uncertainties.

### B.4. Power-moderated mean

The M–P mean does not counteract possible errors in the relative uncertainties when the data set appears to be consistent, this is when \( \hat{x}_{\text{obs}} \) is not (much) larger than unity. Thus data with understated uncertainty have a negative effect on the robustness and the calculated uncertainty. On the contrary the PMM [4, 5] allows moderating the relative weighting also for seemingly consistent data sets.

For the M–P mean as well as for the classical weighted mean, uncertainties \( u_i \) are used with a power of 2 in the weighting factor. By lowering this power, the influence of understated uncertainties can be moderated. A smooth transition from weighted to arithmetic mean can be realized by intermixing the uncertainties associated with both.

Like with the M–P mean, the variances are increased by an unexplained amount \( s^2 \) to ascertain that \( \hat{x}_{\text{obs}} \) is not larger than unity. Then, a variance per datum is calculated for an unweighted mean, taking the larger value between the sample variance and the combined augmented uncertainties:

\[
S^2 = N \max \left[ \sum_{i=1}^{N} \frac{1}{u_i^2 + s^2} \right]^{-1}, \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N(N - 1)}.
\]  

(B.11)

In the expressions (B.10), the weighting factor \( 1/(u_i^2 + s^2) \) is replaced by

\[
\left[ \left( \sqrt{u_i^2 + s^2} \right)^{\alpha} S^{2-\alpha} \right]^{-1},
\]  

(B.12)

in which the power \( \alpha (0 \leq \alpha \leq 2) \) is the leverage by which the mean can be smoothly varied between arithmetic mean (\( \alpha = 0 \)) and M–P mean (\( \alpha = 2 \)). The M–P method can be regarded as a subset of the PMM. Reducing \( \alpha \) has a similar effect on the KCRV and its uncertainty as that obtained by increasing the laboratory variances in the M–P method.

The choice of \( \alpha \) can be made to reflect the level of trust in the stated uncertainties. For data sets with a predominance of understated uncertainties, one obtains a more realistic uncertainty on the KCRV by reducing the power \( \alpha \). This is particularly recommended with small data sets. Larger data sets have a better defined \( \hat{x}_{\text{obs}} \), facilitating the identification of extreme data and the level of reliability of the \( u_i \). As a practical rule, for data sets in which the uncertainties \( u_i \) are informative but frequently understated, one can make the power \( \alpha \) depend on the number of data \( N \) via a heuristic formula

\[
\alpha = 2 - \frac{3}{N}.
\]  

(B.13)

### B.5. DerSimonian–Laird estimator

Owing to its popularity and simplicity, the D–L estimator [21–23] has been included in the simulation study in annex A. Similar to the M–P mean, it increases the uncertainties by a common amount and applies equations (B.10) to calculate \( x_{\text{ref}} \) and \( u(x_{\text{ref}}) \). An appealing feature is that the added variance \( s^2 \) is not derived from an iterative procedure, but directly from the data scatter around the weighted mean \( x_{\text{WM}} \):

\[
\tau^2 = \max \{0, \frac{Q - (N - 1)}{S_1 - S_2/S_1} \},
\]  

(B.14)

where

\[
Q = \hat{x}_{\text{obs}}^2(N - 1) - \sum_{i=1}^{N} \frac{(x_i - x_{\text{WM}})^2}{u_i^2},
\]  

(B.15)

and \( S_1 = \sum u_i^2, S_2 = \sum u_i^{-4} \).

Whereas the D–L is an efficient estimator of the mean, it tends to underestimate the uncertainty of the D–L mean for data sets with predominantly understated uncertainties. The robustness of the D–L mean can in principle be improved by implementing the principles of power moderation (equations (3)–(6) and outlier identification (equation (7))) in a similar way as was done for the M–P mean.

### Annex C. Formulae for ‘degree of equivalence’ and ‘outlier identification’

#### C.1. Outlier identification

For a consistent data set with reliable standard uncertainties, one could apply a weighted mean and use the reciprocal of the variances as weighting factors. Data not complying with this consistent set can be recognized if their difference \( e_i \) from the mean exceeds the associated uncertainty by a factor \( k \) or more.
Table 1. Estimators of mean.

| Estimator $x_{ref} = \sum_{j=1}^{N} w_j x_j$ | Normalized weight $w_i$ | Mean $x_{ref}$ | Associated standard uncertainty $u(x_{ref})$ |
|------------------------------------------|------------------------|----------------|-----------------------------------|
| AM                                      | $1/N$                  | $1/N \sum_{j=1}^{N} x_j$ | $\left[ \sum_{i=1}^{N} (x_i - \bar{x})^2/N(N-1) \right]^{1/2}$ |
| WM                                      | $u^2(x_{ref})/u_i^2$   | $u^2(x_{ref}) \sum_{j=1}^{N} x_j/u_i^2$ | $\left[ \sum_{i=1}^{N} u_i^2 \right]^{1/2}$ |
| M–P, D–L                                | $u^2(x_{ref}) + x^2$   | $u^2(x_{ref}) \sum_{j=1}^{N} x_j/u_i^2 + x^2$ | $\left[ \sum_{i=1}^{N} u_i^2 + x^2 \right]^{1/2}$ |
| PMM                                     | $u^2(x_{ref}) [(\sqrt{\sum_{j=1}^{N} x_j^2})/N]^{-2/α}$ | $u^2(x_{ref}) \sum_{j=1}^{N} x_j/ [\sqrt{\sum_{j=1}^{N} x_j^2}]^{2/α}$ | $\left[ \sum_{i=1}^{N} (u_i^2 + x^2)^{-α} \right]^{1/2}$ |

$u^2(e_i) = (1 - 2w_i)u_i^2 + u^2(x_{ref})$ \ ($x_i$ incl. in mean),

$= (1 - 2w_i)\frac{u^2(x_{ref})}{w_i} + u^2(x_{ref})$ \ (weighted mean),

$= u^2(x_{ref}) \left[ \frac{1}{w_i} - 1 \right]$ \ (weighted mean). \ (C.1)

A similar equation, with opposite sign, holds for data not included in the mean.

$u^2(e_i) = u_i^2 + u^2(x_{ref})$ \ ($x_i$ excl. from mean),

$= u^2(x_{ref}) \left[ \frac{1}{w_i} + 1 \right]$ \ (weighted mean). \ (C.2)

Typical intercomparison data contain understated uncertainties, and normal criteria for outlier identification would reject many data as extreme. Preferably, the identification and rejection of extreme data is kept to a minimum, so that the mean is based on a large subset of the available data. This is easily achieved in the philosophy of the M–P mean, even for discrepant data sets, by using the augmented uncertainties

$u^2(e_i) = (1 - 2w_i)(u_i^2 + x^2) + u^2(x_{ref})$ \ (M–P mean). \ (C.3)

Similarly, one can apply the power-moderated uncertainty for the PMM. In all cases, the variance equations reduce to the same elegant solutions as in equations (C.1) and (C.2), expressed as a function of the normalized weighting factor. If the method reduces to an arithmetic mean ($α = 0$), the weighting factors are equal for all data, irrespective of the stated uncertainty. Extreme data are then identified from their difference with the mean relative to a typical uncertainty per datum.

C.2. Degrees of equivalence

The degrees of equivalence between pairs of NMIIs are not influenced by the estimator used for the KCRV. The degree of equivalence of laboratory data with respect to the KCRV involves calculation of the difference

$$d_i = x_i - x_{ref} = x_i - \sum_{j=1}^{N} w_j x_j,$$ \ (C.4)

and its expanded uncertainty [18]. In the expression C.4, the factor $w_i$ is

• the normalized weighting factor (see table 1) for data included in the mean
• zero for data excluded from the calculation of the KCRV

Data excluded from the calculation of the mean are not correlated with it and the variance associated with $d_i$ is, in this case, the sum of two variances:

$$u^2(d_i) = u_i^2 + u^2(x_{ref})$$ \ ($x_i$ excl. from mean). \ (C.5)

The data that have been included in the calculation of the mean are correlated with it, and the variance of the difference is calculated from

$$u^2(d_i) = (1 - w_i)^2 u_i^2 + \sum_{j=1}^{N} w_j^2 u_j^2$$ \ ($x_i$ incl. in mean),

$$= (1 - 2w_i)u_i^2 + u^2(x_{ref})$$ \ ($x_i$ incl. in mean). \ (C.6)

C.3. Exclusion versus inclusion

Exclusion of a datum $x_i$ from the KCRV is neutral to the DoE of the corresponding laboratory $i$, only under the condition that the (unnormalized) weighting factors assigned to the other laboratories remain unaltered. For the PMM, this is valid if the power $α$ and the added variance $x^2$ are unchanged. These conditions are fulfilled with a consistent data set ($x^2 = 0$) having a large number of data ($α = 2–3/N ≈ 2$) or using a constant $α$.

In practice, the statistical criteria are such that outliers tend to make the data set discrepant. Removal of those data then reduces the $x^2$ value needed to reach $X^2_{obs} = 1$. This results in a better KCRV with smaller uncertainty, but also in DoEs that are higher for the outlier(s) and generally more realistic for the remaining consistent data.

References

[1] Chunovkina A G, Elster C, Lira I and Wöger W 2008 Analysis of key comparison data and laboratory biases Metrologia \bf{45} 211–6
[2] Shirono K, Tanaka H and Ehara K 2010 Bayesian statistics for determination of the reference value and degree of equivalence of inconsistent comparison data Metrologia 47 444–52

[3] Pommé S 2006 An intuitive visualisation of intercomparison results applied to the KCDB Appl. Radiat. Isot. 64 1158–62

[4] Pommé S 2012 Determination of a reference value, associated standard uncertainty and degrees of equivalence—for CCRI(II) key comparison data JRC Scientific and Policy Reports EUR 25355 EN (EUR—Scientific and Technical Research Series) (doi:10.2787/61338 2012)

[5] Pommé S and Spasova Y 2008 A practical procedure for assigning a reference value and uncertainty in the frame of an interlaboratory comparison Accred. Q. Assur. 13 83–9

[6] Mandel J and Paule R C 1970 Interlaboratory evaluation of material with unequal number of replicates Anal. Chem. 42 1194–7

[7] Paule R C and Mandel J 1982 Consensus values and weighting factors J. Res. Natl Bur. Stand. 87 377–85

[8] Kacker R, Datla R and Parr A 2002 Combined result and associated uncertainty from interlaboratory evaluations based on the ISO Guide Metrol. 39 279–93

[9] Birge R T 1932 The calculation of errors by the method of least squares Phys. Rev. 40 207–27

[10] Steele A G and Douglas R J 2006 Extending chi-squared statistics for key comparisons in metrology J. Comput. Appl. Math. 192 51–8

[11] Beissner K 2002 On a measure of consistency in comparison measurements Metrologia 39 59–63

[12] Steele A G, Wood B M and Douglas R J 2005 Outlier rejection for the weighted-mean KCRV Metrologia 42 32–8

[13] Cox M 2007 The evaluation of key comparison data: determining the largest consistent subset Metrologia 44 187–200

[14] Lira I 2007 Combining inconsistent data from interlaboratory comparisons Metrologia 44 415–21

[15] Rukhin A L 2009 Weighted means statistics in interlaboratory studies Metrologia 46 323–31

[16] Duewer D 2008 A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers Accred. Q. Assur. 13 193–216

[17] Pann U and Bremer W 2009 Considerations with regard to the calculation of key comparison reference values and uncertainties Accred. Q. Assur. 14 281–7

[18] Ratel G 2005 Evaluation of the uncertainty of the degree of equivalence Metrologia 42 140–4

[19] Bailat C J et al 2014 International comparison CCRI(II)-S7 on the analysis of uncertainty budgets for 4πβγ coincidence counting Metrologia 51 06018

[20] Spasova Y, Pommé S and Wätjen U 2007 Visualisation of interlaboratory comparison results in PomPlots Accred. Q. Assur. 12 623–7

[21] DerSimonian R and Laird N 1986 Meta-analysis in clinical trials Control. Clin. Trials 7 177–88

[22] DerSimonian R and Kacker R 2007 Random-effects model for meta-analysis of clinical trials: an update Contemp. Clin. Trials 28 105–14

[23] Jackson D, Bowden J and Baker R 2010 How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? J. Stat. Plan. Inference 140 961–70

[24] Ratel G 2006 Median and weighted median as estimators for the key comparison reference value (KCRV) Metrologia 43 S244–8

[25] Michotte C, Pearce A K, Cox M G and Gostely J-J 2006 An approach based on the SIR measurement model for determining the ionization chamber efficiency curves, and a study of 65Zn and 201Tl photon emission intensities Appl. Radiat. Isot. 64 1147–55