CCAUV/02-36

DRAFT FOR DISCUSSION
Proposed guidelines for the evaluation of key comparison data

W Bich[*] M Cox[†] T Estler[‡] L Nielsen[§] W Woeger[¶]

BIPM Director's Advisory Group on Uncertainties

16 April 2002

Two procedures are proposed for the statistical analysis of key comparison measurements. They apply to the simple circulation of a single travelling standard around all the participants. The application of the procedures to a specific set of key comparison measurements provides a key comparison reference value and its uncertainty, the degree of equivalence of the measurement made by each participating national institute and the degrees of equivalence between measurements made by all pairs of participating institutes. Procedure A is based on the use of the weighted mean, together with consistency checks based on classical statistics regarding its applicability. Should the checks fail, action to remedy the situation is suggested. If the remedy is inappropriate, Procedure B can be applied instead. It is based on the use of the median as a more robust estimator in the circumstances.

It is hoped that, following review, web-based analysis software can be provided to support these guidelines.

1 Introduction

The two procedures given here provide implementations of the definitions given in the technical supplement to the Mutual Recognition Arrangement (MRA) [1]:

1. The degree of equivalence of each national measurement standard is expressed quantitatively by two terms:
   (a) Its deviation from the key comparison reference value.
   (b) The uncertainty of this deviation at the 95% level of confidence.
2. The degree of equivalence between pairs of national measurement standards is expressed quantitatively by two terms:
   (a) The difference of their deviations from the key comparison reference value.
   (b) The uncertainty of this difference at the 95% level of confidence.

"National measurement standard" is interpreted as the result of the measurement made by the respective participating national institute of a travelling standard.

Both procedures apply when the measurements relate to a stable travelling standard (Condition 1 in Section 2) and when the measurement of each institute is realised independently of the

[*]Istituto di Metrologia "G. Colonnetti" (IMGC), Italy
[†]National Physical Laboratory (NPL), UK
[‡]National Institute of Standards and Technology (NIST), USA
[§]Danish Institute of Fundamental Metrology (DFM), Denmark
[¶]Physikalisch-Technische Bundesanstalt (PTB), Germany

measurements of other institutes (Condition 2 in Section 2). If, for each institute, a Gaussian distribution can be assigned to the measurand of which the institute's measurement is an estimate (Condition 3 in Section 2), Procedure A should be applied, at least initially. Procedure B can then be used instead if a consistency check made within Procedure A fails. It can also be used *a priori* if Condition 3 is not applicable to the measurements from one or more institute.

NOTE. If other conditions apply, e.g.,

- Some or all of the institutes' measurements are mutually dependent
- The travelling standard is not stable
- A pattern for the comparison is adopted that is different from the simple circulation of a single travelling standard around all the participants
- The key comparison reference value is provided in advance by some means
- A number of travelling standards are circulated and are to be treated together
- Each participant measures the travelling standard at each of a number of stipulated values of a parameter such as wavelength or frequency,

the procedures here may not be valid without appropriate modification. BIPM Director's Advisory Group on Uncertainties intends to develop further guidelines to cover these and other circumstances. It also intends to provide guidelines for linking the key comparisons carried out under the auspices of the CIPM and those operated by regional metrology organisations.

2 Conditions of use

Procedure A is applicable to key comparisons where the following three conditions apply:

1. Each participating national institute provides a measurement of a travelling standard having good short-term stability and stability during transport [1, Appendix F], and the associated standard uncertainty.
2. Each institute's measurement is realised independently of the other institutes' measurements in the key comparison.

   NOTE. The implication of this condition is that there is no mutual dependence of the institute's measurements.
3. For each institute a Gaussian distribution (with mean equal to the institute's measurement and standard deviation equal to the provided standard uncertainty) can be assigned to the measurand of which the institute's measurement is an estimate.

Procedure B is appropriate when the first two conditions apply, but some of the institutes' measurements are inconsistent with the remainder and cannot be removed or corrected.

The use of Procedure A is encouraged when it is applicable.

3 Rationale

The key comparison reference value is interpreted as an estimate of the measurand on the basis of the measurements provided by the participating laboratories.

The weighted mean of the institutes' measurements, where the weights are equal to the reciprocals of the squares of the associated standard uncertainties, should generally be taken as the key comparison reference value [5]. It would, however, be inappropriate if some of the institutes' measurements were inconsistent with the remainder. In such a situation possible reasons for the inconsistency need to be investigated. If time permits, discussions with the relevant laboratories and investigations by those laboratories should take place. The result of the discussions and investigations would hopefully provide corrected measurements for those judged discrepant. In some (ideally very few) cases, such as when the resolution of a difference of opinion is not forthcoming, it may be appropriate to remove discrepant measurements and the analysis repeated with the remaining measurements.

NOTE 1. The pilot institute is responsible for the preparation of a report on the comparison. The report passes through a number of stages before publication. The first draft, draft A, is prepared as soon as all the results have been received from the participants. It includes the results transmitted by the participants, identified by name. It is confidential to the participants [1, Appendix F].

NOTE 2. Once all participants have been informed of the results, individual values and uncertainties may be changed or removed, or the complete comparison abandoned, only with the agreement of all participants and on the basis of a clear failure of the travelling standard or some other phenomenon that renders the comparison or part of it invalid [1, Appendix F]. Thus, the only realistic opportunity for changing or removing the data initially supplied is before Draft A has been distributed to the participants.

Even if permitted, the correction or removal of measurements may be inappropriate for a variety of reasons. Further, for such instances, there may be key comparisons where insufficient time or effort is available to modify the measurements in a scientifically informed way. In these cases it will be necessary to adopt a form of robust analysis that by definition would not be statistically as meaningful as the use of the weighted mean for consistent measurements. If such an analysis were sufficiently robust, in that it exhibited resilience to discrepant measurements, it would be expected to provide a more suitable result for measurements containing discrepancies than would be provided by the weighted mean. Such an analysis is based on the use of the median as the key comparison reference value.

In the case where the weighted mean is taken, the definitions in Section 1 can readily be implemented using a least-squares approach [4].

The implementation is not as straightforward when the median is used as the key comparison reference value, since classical theory is no longer applicable. A method based on propagating the Gaussian distributions assigned as in Condition 3 of Section 2 to evaluate the uncertainty of the key comparison reference value and to provide the degrees of equivalence is proposed in this case. The method can also be used when distributions other than Gaussian are assigned. See Appendix A.

4 The input quantities to the analysis

Identify the participating institutes, N in all, by the numbers $i = 1, \ldots, N$. The input quantities to the analysis are the institutes' measurements, denoted by x_i, $i = 1, \ldots, N$, and the standard uncertainties of these values, denoted by $u(x_i)$, $i = 1, \ldots, N$.

NOTE. If, on examination of the complete set of results, the pilot institute finds results that appear to be anomalous, the corresponding institutes are invited to check their results for numerical errors but without being informed as to the magnitude or sign of the apparent anomaly. If no numerical error is found the result stands and the complete set of results is sent to all participants [1, Appendix F].

5 Procedure A

This section contains the recommended procedure for the analysis of key comparison measurements when all three conditions of Section 2 apply. It is based on the use of least-squares adjustment.

1. Determine the weighted mean y of the institutes' measurements, using the inverses of the squares of the stated standard uncertainties as the weights:

$$y = \frac{x_1/u^2(x_1) + \cdots + x_N/u^2(x_N)}{1/u^2(x_1) + \cdots + 1/u^2(x_N)}. \tag{1}$$

NOTE. y so computed is a valid estimate if the provided measurements are consistent. See Step 3.

2. Determine the standard deviation $u(y)$ of y from

$$\frac{1}{u^2(y)} = \frac{1}{u^2(x_1)} + \cdots + \frac{1}{u^2(x_N)}. \tag{2}$$

3. Apply a chi-squared test to carry out an overall consistency check of the results obtained [4]:

   (a) Form the observed chi-squared value

   $$\chi^2_{\text{obs}} = \frac{(x_1 - y)^2}{u^2(x_1)} + \cdots + \frac{(x_N - y)^2}{u^2(x_N)}.$$

   (b) Assign the degrees of freedom

   $$\nu = N - 1.$$

   (c) Regard the consistency check as failing if

   $$\Pr\left\{\chi^2(\nu) > \chi^2_{\text{obs}}\right\} < 0.05.$$

   NOTE 1. "Pr" denotes "probability of".

   NOTE 2. This test assumes normality, and therefore depends on Condition 3 in Section 2.

4. If the consistency check does not fail:

   (a) Accept y as the key comparison reference value x_{ref}.

   (b) Accept $u(y)$ as the standard uncertainty $u(x_{\text{ref}})$ of the key comparison reference value.

   (c) Calculate the degrees of equivalence:

      i. For $i = 1, \ldots, N$ form the degree of equivalence of institute i as the pair of values $(d_i, U(d_i))$ using

      $$d_i = x_i - x_{\text{ref}}, \tag{3}$$

      $$U(d_i) = 2u(d_i), \tag{4}$$

      where $u(d_i)$ is given by

      $$u^2(d_i) = u^2(x_i) - u^2(x_{\text{ref}}). \tag{5}$$

      NOTE 1. The factor 2 in Formula (4) and elsewhere gives 95% coverage under the assumption of normality (Condition 3 in Section 2.)

      NOTE 2. The formula for $u^2(d_i)$ involves a *difference* of two variances as a consequence of the mutual dependence of x_i and x_{ref}. It is established in Appendix C.

      ii. For $i = 1, \ldots, N$ and $j = 1, \ldots, N$, with $j \neq i$, form the degree of equivalence between institute i and institute j as the pair of values $(d_{i,j}, U(d_{i,j}))$ using

      $$d_{i,j} = x_i - x_j,$$

      $$U(d_{i,j}) = 2u(d_{i,j}),$$

      where $u(d_{i,j})$ is given by

      $$u^2(d_{i,j}) = u^2(x_i) + u^2(x_j).$$

   NOTE 1. The difference $d_{i,j}$ of the deviations of institute measurements x_i and x_j from the key comparison reference value x_{ref} does not depend on x_{ref} since, using Definition 2a in Section 1,

   $$d_{i,j} = d_i - d_j = (x_i - x_{\text{ref}}) - (x_j - x_{\text{ref}}) = x_i - x_j.$$

   NOTE 2. The formulae for $U(d_i)$ and $U(d_{i,j})$ are based on the Gaussian distributions for the measurands of which these quantities are estimates. These Gaussian distributions follow from Condition 3 in Section 2.

(d) Record the results obtained and the manner in which they were determined.

(e) Finish.

5. (The consistency check has failed.) If (a) Draft A has not been provided to partcipants, (b) adequate time is available and (c) it is economically viable to do so, investigate the reasons for the inconsistency:

   (a) An investigation would involve:

      i. Identify discrepant measurements [4]: if

      $$|d_i| > 2u(d_i),$$

      classify x_i as discrepant at the 5% level of significance.

      NOTE 1. This test assumes normality, and therefore depends on Condition 3 in Section 2.

      NOTE 2. On the basis of *statistical variability alone*, 5% of measurements would be expected to be classified as discrepant.

      ii. Discuss the matter with the laboratories concerned, obtaining corrected measurements and uncertainties where appropriate.

      iii. If all laboratories concerned provide corrected measurements, return to Step 1.

      iv. (Not all laboratories concerned provided corrected measurements and uncertainties.) If all laboratories concerned are prepared to withdraw from the comparison, delete the measurements for those laboratories and return to Step 1, after relabelling the input quantities (Section 4) appropriately.

6. If this point is reached, one of the following reasons applies: (a) Draft A has been provided to participants, (b) the situation is not resolved, i.e., at least one institute is regarded as providing a discrepant measurement and it is unprepared to withdraw from the comparison, or (c) neither adequate time is available nor is it economically viable to investigate the reasons for the inconsistency. The alternative procedure in Section 6 can be used.

6 Procedure B

This section contains an alternative procedure, based on the use of the median (or some other suitable estimator) as the key comparison reference value for the analysis of key comparison measurements. It can be applied when the procedure of Section 5 is inappropriate. Thus, it can be used when Step 6 of Procedure A is reached, or *a priori* when Condition 3 of Section 2 does not apply.

NOTE. The median can be considered appropriate when

1. The measurement by each institute can be regarded as equally likely to lie above or below the required reference value, and the provided uncertainties are to be disregraded for establishing the reference value, but
2. The provided uncertainties are to be utilised in evaluating the uncertainties associated with the reference value and the degrees of equivalence.

1. For each input quantity:

   (a) If the *only* information available is the measurement and its standard uncertainty, assign a Gaussian distribution to that input quantity, in accordance with Condition 3 of Section 2.

   (b) If other information is available for the input quantity, assign the probability distribution to that input quantity that is (minimally) consistent with this knowledge.

   NOTE. Bayes' Theorem or the Principle of Maximum Entropy may be applied for this purpose.

2. Decide the choice of statistical estimator to be used for the key comparison reference value. Here, the median is used as a possible robust estimator, but an alternative estimator can be used in its place if justification is provided. Simply replace all occurrences below of the term "median" by the name of the alternative estimator.

   NOTE 1. The median can be expected to be more appropriate than the weighted mean if a number (up to one third, say) of the institutes' measurements can be regarded as discrepant.

   NOTE 2. The median has been used in some key comparisons, as has the weighted mean.

3. Decide a large number M of Monte Carlo trials.

   NOTE. $M = 10^6$ is recommended.

4. For $r = 1, \ldots, M$:

   (a) Sample at random from the probability distribution for each of the input quantities to obtain the column vector

   $$\mathbf{x}^{(r)} = (x_1^{(r)}, \ldots, x_N^{(r)})^{\mathrm{T}}.$$

   NOTE 1. Advice on sampling from probability distributions is available [2].

   NOTE 2. This sample of the input quantities is, according to the assigned probability distributions, as legitimate as any other such sample and hence as legitimate as the provided institutes' measurements.

   (b) Form the median $m^{(r)}$ of this sample.

5. Assemble the M column vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$ into an $N \times M$ matrix Z:

   $$Z = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}).$$

6. Form the row vector

   $$\mathbf{q} = (m^{(1)}, \ldots, m^{(M)}),$$

   regarding the M values in $\mathbf{q}$ as describing the probability distribution of the median estimator of the key comparison reference value.

7. Take the mean of the values in $\mathbf{q}$ as the key comparison reference value x_{ref}.

8. Take the standard deviation of the values in $\mathbf{q}$ as the standard uncertainty $u(x_{\mathrm{ref}})$ of x_{ref}.

9. Use $\mathbf{q}$ in the manner described in Appendix B to form the shortest coverage interval at the 95% level of confidence for the measurand of which x_{ref} is an estimate.

10. Calculate the degrees of equivalence:

    (a) For $i = 1, \ldots, N$ form the degree of equivalence of institute i as the pair of values $(d_i, U(d_i))$ as follows:

    i. Form $d_i = x_i - x_{\mathrm{ref}}$.

    ii. Form the row vector $\mathbf{r}$ given by

    $$\mathbf{r} = (\text{row } i \text{ of } Z) - \mathbf{q}.$$

    iii. Regard $\mathbf{r}$ as describing the probability distribution of D_i, the measurand of which d_i is an estimate.

    iv. Use $\mathbf{r}$ in the manner described in Appendix B to form the shortest coverage interval at the 95% level of confidence for D_i.

    (b) For $i = 1, \ldots, N$ and $j = 1, \ldots, N$, with $j \neq i$:

i. Form the row vector $\mathbf{r}_{i,j}$ given by

$$\mathbf{r}_{i,j} = (\text{row } i \text{ of } Z) - (\text{row } j \text{ of } Z).$$

ii. Regard $\mathbf{r}_{i,j}$ as describing the probability distribution of $D_{i,j}$, the measurand of which $d_{i,j}$ is an estimate.

iii. Use $\mathbf{r}_{i,j}$ in the manner described in Appendix B to form the shortest coverage interval at the 95% level of confidence for $D_{i,j}$.

(c) Finish.

11. Record the results obtained and the manner in which they were determined.

NOTE 1. Steps 2–11 can be applied with "median" replaced by "weighted mean". That application would serve as a further validation for the case of consistent measurements. The results obtained would then be expected to be identical to those that would be produced by the main procedure, apart from the effects of sampling from the probability distributions.

NOTE 2. The computations of Steps 4–10 can be expected to take a few seconds on a PC operating at 1 GHz or faster.

Acknowledgements

Peter Harris of the National Physical Laboratory made a substantial input to the preparation of these guidelines.

Some 40 scientists and metrologists at a number of national measurement institutes reviewed a draft of these guidelines, as did several members of CIPM Consultative Committees.

References

[1] BIPM. Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes. Technical report, Bureau International des Poids et Mesures, Sèvres, France, 1999.

[2] M. G. Cox, M. P. Dainton, and P. M. Harris. Software specifications for uncertainty evaluation and associated statistical analysis. Technical Report CMSC 10/01, National Physical Laboratory, Teddington, UK, 2001. Available at http://www.npl.co.uk/ssfm/index.html.

[3] JCGM. Guide to the Expression of Uncertainty in Measurement. Supplement 1: Numerical Methods for Propagating Probability Distributions. Technical report, Joint Committee for Guides in Metrology, 2002. Draft.

[4] L. Nielsen. Evaluation of measurement intercomparisons by the method of least squares. Technical Report DFM-99-R39, Danish Institute of Fundamental Metrology, Denmark, 2000.

[5] K. Weise and W. Wöger. Removing model and data non-conformity in measurement evaluation. *Measurement Sci. Technol.*, 11:1649–1658, 2000.

Appendices

A The use of the median (or some other estimator) as a key comparison reference value

In this appendix the use of the median (or some other estimator) as a key comparison reference value is considered. Once an estimate of the key comparison reference value has been determined, it is straightforward to form the deviation of each institute's measurement from the key comparison reference value (Definition 1a in Section 1) and the difference of the deviations of two institute measurements from the key comparison reference value (Definition 2a in Section 1).

It is not in general as straightforward to calculate the associated uncertainties as it is for the weighted mean. An approach based on the principle of the propagation of distributions [3] can be used. Consider three simple models:

1. The formula for the median. It relates the institutes' measurements as input quantities to the median as measurand.

   NOTE. In place of "median", the name of any other estimator can be substituted.

2. The formula for the deviation of an institute's measurement from the key comparison reference value, with the institutes' measurement and the key comparison reference value as input quantities and the deviation as measurand.

3. The formula for the difference of the deviations of two institutes' measurements from the key comparison reference value, with the institutes' measurements as input quantities and the difference as measurand.

If Condition 3 in Section 1 applies, assign a Gaussian distribution with mean equal to the institute's measurement and standard deviation equal to the provided standard uncertainty to the measurand of which the institute's measurement is an estimate. Otherwise, assign the probability distribution that is appropriate. By propagating these distributions through the models, the required uncertainties can be evaluated. See Section 6.

NOTE. If information additional to the mean and standard deviation of the distribution is available, Bayes' theorem or the Principle of Maximum Entropy may be useful in assigning an appropriate distribution.

B Determination of coverage intervals

A coverage interval for a measurand Y at the 95% level of confidence (or at some other level) can be determined from the distribution function $G(Y)$ of Y. The coverage interval is not generally unique.

In particular, the endpoints of a 95% coverage interval for the measurand are given by the 0.025- and 0.975-fractiles of $G(Y)$, i.e., the values of y given by $G^{-1}(0.025)$ and $G^{-1}(0.975)$.

If Monte Carlo Simulation has been used, as in the procedure in Section 6, a 95% coverage interval can be determined as follows:

1. Denote by $y_1, \ldots, y_M$ the M values of Y obtained by Monte Carlo Simulation.
2. Sort these values into non-decreasing order, denoting the sorted values by $y_{(1)}, \ldots, y_{(M)}$.
3. Form a 95% coverage interval $(y_{(\lfloor 0.025M \rfloor)}, y_{(\lceil 0.975M \rceil)})$, where $\lfloor v \rfloor$ is the largest integer no greater than v and $\lceil v \rceil$ is the smallest integer no smaller than v.

The coverage interval so obtained is *central* with respect to probability, i.e., 2.5% of the distribution of possible values lies to the left of the interval and 2.5% to the right. It will not generally be the shortest coverage interval, unless the distribution is symmetric. In particular, the use of the median will typically give rise to an asymmetric distribution.

The most general 95% coverage interval is given by the p– and $(p + 0.95)$–fractiles of $G(Y)$, with $0 \leq p \leq 0.05$.

NOTE. The choice $p = 0.025$ is natural for a $G(Y)$ corresponding to a symmetric distribution. It has the shortest length in this case.

The shortest interval can be determined numerically by taking a sequence of closely spaced values of probability p between zero and 0.05. For each such value the length of the coverage interval whose endpoints correspond to p and $p + 0.95$ is computed. The shortest of these intervals is then taken. So, to determine the shortest interval in general, in place of Step 3 above:

3a. Define the inverse $\widehat{G}^{-1}(p)$ of the empirical distribution function as the piecewise-linear function joining the points $(p_r, y_{(r)})$, $r = 1, \ldots, M$, where $p_r = (r - 1/2)/M$.

NOTE. p_r is the rth in a sequence of uniformly-spaced probability values centred on M contiguous probability intervals of width $1/M$.

3b. For $r = 1, \ldots, M$:

1. Set the rth in a sequence of M uniformly spaced probability values between p_1 and $p_M - 0.95$:
$$\rho_r = \frac{1}{2M} + \left(\frac{1}{M} - \frac{0.95}{M-1}\right)(r - 1).$$
2. Form the length
$$L_r = \widehat{G}^{-1}(\rho_r + 0.95) - \widehat{G}^{-1}(\rho_r)$$
of the 95% coverage interval
$$(\widehat{G}^{-1}(\rho_r), \widehat{G}^{-1}(\rho_r + 0.95)).$$

NOTE. Linear interpolation at ρ_r and $\rho_r + 0.95$ of the points $(p_r, y_{(r)})$, $r = 1, \ldots, M$, provides the required values.

3c. Take as the shortest coverage interval the interval
$$(\widehat{G}^{-1}(\rho_s), \widehat{G}^{-1}(\rho_s + 0.95)).$$
where s is a value such that
$$L_s \le L_r, \qquad r = 1, \ldots, M.$$

NOTE 1. The shortest coverage interval may not be unique.

NOTE 2. Because the coverage intervals are obtained approximately, there may be several having lengths that are close to the length of the shortest so obtained. Therefore, a choice can be made, taking account of this consideration.

C The uncertainty of the degree of equivalence of an institute

This appendix establishes the result (5) for the standard uncertainty of the degree of equivalence of institute i.

Define
$$\omega_i = \frac{u^2(x_{\text{ref}})}{u^2(x_i)}. \tag{6}$$

Then, from (1), (2) and $x_{\text{ref}} \equiv y$,
$$x_{\text{ref}} = \sum_{i=1}^{N} \omega_i x_i. \tag{7}$$

Note that, using (2),
$$\sum_{i=1}^{N} \omega_i = 1. \tag{8}$$

Then, from (3) and (7)

$$d_i = x_i - x_{\text{ref}} = x_i - \sum_{j=1}^{N} \omega_j x_j = (1 - \omega_i) x_i - \sum_{\substack{j=1 \\ j \neq i}}^{N} \omega_j x_j.$$

So, since there is no mutual dependence in the measurements x_j, $j = 1, \ldots, N$ (Condition 2 of Section 2),

$$u^2(d_i) = (1 - \omega_i)^2 u^2(x_i) + \sum_{\substack{j=1 \\ j \neq i}}^{N} \omega_j^2 u^2(x_j) = ((1 - \omega_i)^2 - \omega_i^2) u^2(x_i) + \sum_{j=1}^{N} \omega_j^2 u^2(x_j).$$

Using (6),

$$u^2(d_i) = (1 - 2\omega_i) u^2(x_i) + \sum_{j=1}^{N} \omega_j u^2(x_{\text{ref}}).$$

Using (8), and (6) again,

$$u^2(d_i) = u^2(x_i) - 2u^2(x_{\text{ref}}) + u^2(x_{\text{ref}}) \sum_{j=1}^{N} \omega_j = u^2(x_i) - u^2(x_{\text{ref}}),$$

which establishes (5).