**CCT/01-41**

# Confidence levels and confidence intervals for key comparisons to facilitate MRA decisions

*A.G. Steele, B.M. Wood, R.J. Douglas*

**Abstract**
Confidence interval methods are presented as important tools for facilitating and describing decisions based on comparison data. We illustrate their utility for associating an acceptance criterion with its corresponding assured confidence level. In this paper we analyze comparison data expressed in standard form using the Quantified Demonstrated Equivalence (*QDE*) approach, and also evaluate the probability of agreement within a specified range. The availability and use of new Excel tools are discussed as a practical means of adding rigour to the many decisions needed for MRA Appendices B and C. For acceptance into Appendix C of the MRA, examples are given that illustrate quantifying equivalence demonstrated by comparisons, and confidence imputed to expert opinion.

## 1. Introduction

In this paper, we express inter-laboratory comparison results as confidence intervals, and show how confidence interval methods can complement traditional null hypothesis testing. Confidence intervals can be particularly helpful in making and describing comparison-based decisions recently added to the tasks of international metrology.

We consider some of the new tasks created by the CIPM Mutual Recognition Arrangement [1]. The tasks are based on certain Key Comparisons that have been, and will continue to be, selected and performed. Laboratories measure a circulated artefact, and each reports its measurement and standard uncertainty [2]. The Key Comparison reports are used as evidence of equivalence of measurements made in participating laboratories.

For making decisions with Key Comparison data, null hypothesis testing has some identified deficiencies that can be remedied with confidence interval interpretations such as Quantified Demonstrated Equivalence (*QDE*) [3]. The *QDE* formalism was developed specifically to deal with the realities of inter-laboratory comparisons. The procedure takes as input two measured values, two uncertainties and a confidence level (e.g. 95%). Its output is a single parameter: the interval expected to contain 95% of re-measurements (i.e. repetitions of the full comparison), centred on the end-user expectation of agreement. *QDE* has an established theoretical development, and it is easily calculated for all cases of practical interest. It can be applied equally to bilateral comparisons and to comparisons with the so-called Key Comparison Reference Values (*KCRV*s). Perhaps most importantly, it is a single-parameter interpretation of a comparison that focuses entirely on expressing the demonstrated agreement in a way that can be clearly expressed to non-experts [4,5].

We develop the confidence interval formalism further to describe the probability for agreement, on a specified interval, between any two results in a Key Comparison. This Quantified Demonstrated Confidence (*QDC*) approach takes as input two measured values, two uncertainties and a specified interval, again centred on the end-user expectation of agreement. The output is the probability (confidence level) that a repetition of the comparison would give a difference that falls within the specified interval.

New computational simplifications are presented for calculating both of these expressions of agreement: the confidence interval (for confidence levels other than 95%) and the confidence probability within a specified interval. These single-parameter descriptions are most appropriate when demonstrated agreement is the paramount concern of a discussion: a comparison that reveals a difference of the means that is many times larger than the combined uncertainties can still be used as supporting evidence for equivalence at a sufficiently lower level of accuracy. Thus even an extreme outlier in a Key Comparison, located many of its standard uncertainties away from the *KCRV* or the values of the other participants, can be used rigorously to demonstrate rigorously an associated Appendix C measurement capability claim, provided that this claim has an appropriately larger uncertainty.

For multilateral comparisons with a large number of participants, a spreadsheet ‘toolkit’ has been prepared to simplify and automate the tasks of creating the “full bilateral equivalence array” used to summarize Key Comparison data. It performs all of the confidence level and confidence interval calculations, as well as a number of other statistical functions, and is designed for use from within Microsoft Excel spreadsheets [6]. The toolkit is freely

available by contacting the authors for a copy. It can be customized locally, or used “as is”, to present data in the MRA Excel format and to support acceptance decisions. This paper is not intended as a detailed description of the toolkit, although it uses use the toolkit in describing and illustrating the *QDE* formalism for facilitating MRA decisions.

After presenting the background and computational methods of the analysis, several examples are presented using recent comparison data. Calculations show the confidence intervals for agreement between pairs of measurements at both the 95% and the 68% confidence levels. As well, the use of *QDC* is discussed as an evaluation tool for accepting a given Institute's claimed capability in the Appendix C database.

## 2. Decisions within the Context of the MRA

The CIPM Mutual Recognition Arrangement (MRA), signed in 1999, formally began a process of measurement comparison, evaluation and organized mutual recognition of national measurement standards and calibration capabilities [1]. Results of Key Comparisons are summarized in a growing online database called Appendix B of the MRA. The Key Comparisons are generally published in the open literature and results are made available within the database in support of the recognized calibration and measurement capabilities that are approved for entry in a second part of the database called Appendix C of the MRA.

At present, the databases in Appendices B and C are still largely unpopulated. Thousands of decisions are being made, and each will be revisited over the years as further information becomes available. Pre-agreed criteria and procedures, such as now exist for some aspects of Key Comparisons, could facilitate the decision-making process, but debate continues about interpreting comparison results and about quantifying the degree of equivalence. The accumulated intra-laboratory experience of metrologists is being extended to deal rigorously with these formalized inter-laboratory comparisons. But the intra-laboratory practices of data pooling and outlier rejection, which have been a routine but informal part of self-evaluation, are not rigorously established as being applicable to the processes of choosing an inter-laboratory mean and excluding some laboratories from Appendix C.

For the MRA, the degree of equivalence between two measurement standards is expressed quantitatively by two parameters that are determined experimentally in a Key Comparison: the difference between the measured values and the uncertainty of that difference. It is a probabilistic form that uses the *ISO Guide to the Expression of Uncertainty in Measurement* [2] norms for uncertainty. It is the canonical two-parameter form of the bilateral agreement demonstrated either by two laboratories in a Key Comparison, or of one laboratory to a Key Comparison’s reference value.

There can be important but subtle differences in interpretation when pooling inter-laboratory data compared to pooling data within one laboratory. Inter-laboratory comparisons raise new issues not usually addressed in established intra-laboratory practices. Similarly, the development of consensus for standards concerning descriptions of correlations in inter-laboratory measurement comparisons has only just begun.

Perhaps constrained by experience within their own laboratories, metrologists are still expressing a significant variety of opinion concerning the requirements of the MRA, the need for interpretation criteria for the MRA, and the final acceptance into Appendix C of the MRA.

## 3. Key Comparisons and Appendix B: the *KCRV*

The fundamental technical basis for the MRA is found in the international comparisons, organized for each major area of metrology, that explore the equivalence of the principal techniques of measurement realized by different National Metrology Institutes (NMIs). These Key Comparisons are each conducted by a CIPM Consultative Committee (CC), and will usually lead to hundreds of inferred bilateral comparisons between the tens of laboratories participating in a Key Comparison.

For some less critical purposes, a usefully simplified description of a Key Comparison will be used with its “Key Comparison Reference Value” (*KCRV*) to summarize the Key Comparison in terms of the tens of laboratory differences with respect to the *KCRV*. The *KCRV* is expected to be a good, but not necessarily the best, representation of the SI value of the comparison. To date, comparisons have used many simple methods for determining a *KCRV*, including the mean, the median, and the value of the pilot (or another participating) laboratory. The *KCRV* may be averaged equally over all laboratories, or weighted unequally, or restricted to laboratories with independent realizations of the unit. Even with a particular method chosen, the data for some Key Comparisons naturally requires more than one *KCRV*. Furthermore, each of the first round of Key Comparisons is envisaged as being followed by an unending series of similar Key Comparisons, spaced some years apart. The Consultative Committees face a continuing and extensive stream of data to be considered and reconsidered.

The MRA Appendix B database also contains a two-dimensional matrix, providing the bilateral degrees of equivalence between each pair of participating NMIs. This is the most complete summary of the Key Comparison, and includes all pair differences and the corresponding pair uncertainties, calculated by the technical experts. In general, this matrix cannot be calculated using only the information contained in the simple table of degrees of equivalence with respect to the *KCRV*, since there are often correlated components within the various uncertainty budgets.

Many problems can be avoided by the choice of an appropriate statistic or representative value for the *KCRV* and its uncertainty, $u_{\mathrm{KCRV}}$ [7]. Explicitly specifying their definitions and any associated assumptions can simplify the task of interpreting and using these quantities. The Appendix B submission format [8] requires a complete description of both the *KCRV* and $u_{\mathrm{KCRV}}$, and includes a table of equivalence that lists the laboratory differences from the *KCRV*, and the uncertainty of this difference. It is up to the pilot laboratory, in consultation with the Key Comparison participants and perhaps the full Consultative Committee, to determine the appropriate methods for determining the *KCRV* and $u_{\mathrm{KCRV}}$.

## 4. Interpretation of Key Comparison Data

Within a single laboratory, approximations deliver computational simplicity, rules of thumb supplement detailed analysis, and the power of intuition gained through experience is employed. These approaches, while still very useful for completing the intra-laboratory part of the analysis, have not yet demonstrated a corresponding applicability for comparisons between laboratories: neither for inter-laboratory comparison analysis, nor for the decision-making process which may result in a laboratory being excluded from Appendix C.

Easily understood quantitative measures such as confidence are helpful both in this discussion and in describing conclusions to other parties. In particular, quantified expressions of confidence concerning equivalence facilitate communication with the targeted audience for Appendix C, which lies almost wholly outside the community of NMI metrologists. For Appendix C acceptance decisions, quantification is desirable to describe the relative "weights" given to (i) Appendix B Key Comparisons, and (ii) supplementary expert opinion relying on other information not contained in these comparisons.

We believe that the simple probability calculus of *QDE* and careful attention to the 'ownership' of the probability distributions can be used to clarify and quantify discussions of equivalence through the use of the joint confidence for replicating a result within an interval. The joint confidence is an integral over the joint probability distribution describing the two results. In the context of this joint distribution, it can be helpful to remember that each laboratory expresses its own opinion as a marginal probability distribution and that a description of correlations is only convincing when both parties agree. There will be at least two parties' opinions reflected in the construction of the joint distribution, and the pilot laboratory might add a third, and the reader, re-examining the comparison, could become a fourth party (e.g. a fourth-party proposal to add an uncertainty component to fix a 'failed' comparison).

To compare results from a Key Comparison, each laboratory's measured mean value, with its associated uncertainty, is considered as representing a separate (and in the absence of explicit correlation information, independent) probability distribution to be preserved throughout the analysis. *QDE* preserves the measured difference of the mean values, appropriately expanded by the stated uncertainties, rather than expending it on a weak, one-sample, test of the stated uncertainties (as is done in conventional null-hypothesis testing). *QDE* can be determined for any comparison that reports uncertainties following the *ISO Guide to the Expression of Uncertainty in Measurement* [2]. *QDE* can determine and present a confidence interval for agreement based on the two laboratories' general distributions, although they will usually be specified as a difference of the means and a standard uncertainty. The convolution of the two independent distributions, $P_{\mathrm{p}}(z) = P_1(z_1) \otimes P_2(z_2)$, expresses the probability distribution of two values differing by $z$. $P_{\mathrm{p}}(z)$ is centred at the difference in the means of the two distributions, rather than being centred at zero. Symmetric integration of the convolution about zero yields the probability, or confidence $C$, that each laboratory will report the same value within the integration limits of a confidence interval $[-d_{\mathrm{C}}, +d_{\mathrm{C}}]$. This confidence interval is conceptually similar to an expanded uncertainty but is taken symmetrically about the reader's expectation that the values should agree (i.e. $z$=0), rather than about the experimental difference. The confidence interval for a particular confidence $C$ is determined by solving Equation 1 to obtain $d_C$.

$$C\% = \int_{-d_C}^{+d_C} P_p(z)dz \cdot 100\% \qquad (1)$$

In cases where correlations exist, the combination of the two distributions to form $P_{\mathrm{p}}(z)$ is no longer a simple convolution, but the probability calculations are not computationally difficult. For normal distributions, correlations

between the uncertainties of the two laboratories, $u_1$ and $u_2$, can be fully described by the pair uncertainty $u_p$ using the correlation coefficient $r_{12}$: $u_p^2 = (u_1^2 + u_2^2 - 2r_{12}u_1u_2)$.

Statisticians from both the Bayesian and frequencist schools must all believe that Equation 1 holds true: a confidence level is an integration over a probability distribution within a specified interval. Equation 1 may be solved for the confidence level, $C$%, given a particular confidence interval $d_C$. It can also be solved for the confidence interval, $d_C$, given a confidence level $C$%.

### *4.1 Confidence within any Interval*

The confidence interval formalism, developed for *QDE*, can be used [9] to address the conventional question: *"In any specified interval, what level of confidence for agreement has been demonstrated by the results for two laboratories in a comparison?"* Note that this question is concise and seeks an interval centred on the expectation of bilateral agreement between two specified results, and does not require either the pooling of results or the null hypothesis. When applied to the usual case of uncertainty budgets reported as normal distributions, the solution is simply expressed using the error function, erf($x$), defined in Equation 2 and used to evaluate the Gaussian integral.

$$\mathrm{erf}(x) = \frac{2}{\sqrt{p}} \int_0^x e^{-t^2} dt \qquad (2)$$

To construct the probability that one laboratory's claim, represented by $m_1 \pm ku_1$, includes another value, represented by $m_2$ and possibly another uncertainty $u_2$ , we first construct the probability for the difference as a convolution: a normal distribution centered at $m_1$ - $m_2$, with standard deviation $u_p$. As usual, $u_p$ is the pair uncertainty and is made up of both uncertainties: $u_p^2 = u_1^2 + u_2^2 - 2r_{12}\,u_1u_2$ , where $r_{12}$ is the correlation coefficient between the two uncertainties. The probability is the integration from $-ku_1$ to $+ku_1$ of the distribution, and Equation 2 gives the answer by shifting origin and scale.

Using the expanded uncertainty $U_1 = ku_1$ in the calculation makes this expression directly comparable with other statistics which are generally used when summarizing data. Note that the *pair* uncertainty $u_p$ enters the expression as a standard deviation, but the interval $[-U_1, +U_1]$ is based on one laboratory's claimed expanded uncertainty, in this case $ku_1$. Usually, $k$ would be 2, and, for uncertainty budgets which conform to the *ISO Guide* [2], it must be specified explicitly if $k \neq 2$.

The resulting expression for the probability that a repeat of the comparison would have $m_2'$ that falls within $[m_1' - ku_1, m_1' + ku_1]$ is the Quantified Demonstrated Confidence (*QDC*), given in Equation 3.

$$QDC(k) = \frac{1}{2}\left( \mathrm{erf}\left[ \frac{m_1 - m_2 + ku_1}{\sqrt{2}u_p} \right] - \mathrm{erf}\left[ \frac{m_1 - m_2 - ku_1}{\sqrt{2}u_p} \right] \right) \bullet 100\% \qquad (3)$$

In Equation 3, the arguments of the error functions mean that the confidence level depends on the difference in values, the uncertainty of that difference, and the laboratory's uncertainty claim. Note that the confidence level is different for the two laboratories when $u_1 \neq u_2$ ; this is one reason why direct comparisons cannot produce convincing conclusions about the validity of the uncertainty claim of the laboratory having the smaller uncertainty value. Note also that when $u_1 = u_2$ , in the absence of correlations, 84% is the highest confidence that can be demonstrated for agreement within $\pm 2u_1$ .

The discussion so far has explicitly focussed on the quantified demonstrated confidence for agreement between two laboratory values in a Key Comparison, but the same line of reasoning can be used to calculate *QDC* for a laboratory and the *KCRV*: to do so, we simply take $m_2 = KCRV$ and $u_2 = u_{\mathrm{KCRV}}$.

Equation 3 is easily extended to include travel uncertainty in $u_p$ , and to use other intervals (an MRA Appendix C claim, for example $u_1'$, in place of $u_1$ in Equation 3 - although $u_1$ must still be used when calculating $u_p$ for the comparison).

### *4.2 Evaluating Confidence Intervals*

Equation 1 may be solved for a given confidence level, $C$%, to obtain the associated confidence interval, $d_C$. In keeping with the common practice of reporting an expanded uncertainty at the 95% level of confidence (about $k$=2

for normal distributions), the particular solution for $C$ = 95% in Equation 1, $d_{0.95}$, has been investigated in detail [3]. This $d_{0.95}$, the Quantified Demonstrated Equivalence at the 95% confidence level is denoted by $QDE_{0.95}$. It may be calculated with the convenient approximation formula shown in Equation 4, using the following coefficients: $a_o(0.95) = 1.645$; $a_1(0.95) = 0.3295$; $a_2(0.95) = 4.05$.

$$QDE_{C\%} \approx \quad |m_1 - m_2| + \{a_0(C) + a_1(C) \times \exp[-a_2(C) \times (|m_1 - m_2|)/u_p]\} u_p \qquad (4)$$

This expression has been extended for use as an approximation to the general solution of Equation 1 for a wide range of confidence levels $C$. Appendix A contains a table of coefficients suitable for use with Equation 4 to determine confidence intervals for $C$ in the range from 0.995 to 0.40, along with a description of the limitations on the approximation formula and a discussion of techniques for generating more precise solutions.

Of particular interest is the Equation 4 approximation with the parameters for $C$=0.68, which closely corresponds to the standard ($k$=1) uncertainty for a normal distribution. In this case, the coefficients are: $a_0(0.68) = 0.468$; $a_1(0.68) = 0.554$; $a_2(0.68) = 2.669$.

It may prove appropriate to use $QDE_{0.68}$, or $QDE$ for some other confidence level, rather than $QDE_{0.95}$ in assessing Appendix C claims when the confidence demonstrated by a comparison is to be supplemented with other information or expert opinion. Even if 95% confidence is regarded by some NMIs as an appropriate confidence level for approval, they may still wish to use $QDE$ for a lower confidence level and give a quantified weight to other information or expert opinion. Sometimes this is unavoidable: for some Appendix C entries, there will be no directly applicable Key Comparison, and decisions will have to rely on expert opinion and other information. The $C$=99.5% coefficients, used in evaluating the $QDE_{0.995}$ confidence interval, are useful when preparing expanded uncertainty claims of equivalence at $k$=3, such as might be wanted by an instrumentation manufacturer writing performance specifications.

The $QDE_{0.68}$ and $QDE_{0.95}$ quantified demonstrated equivalence intervals are useful in graphical representations of comparison data for making rapid visual assessments of equivalence, in the same fashion that $k$=1 and $k$=2 error bars are normally used. For normal uncertainty distributions, comparison data pairs plotted with either of these two particular $QDE$ "error bars" (derived from that pair difference and pair uncertainty) will always overlap each other's mean value. Notice that $QDE_{C\%}$, like any confidence interval, has the same dimensions as the quantity in the comparison; the confidence level or probability, $C$, is a dimensionless fraction, just like any other probability.

## 5. A Confidence Interval Software Toolkit

Spreadsheet applications are widely used for quick analysis and summary of the results of Key Comparisons performed by NMIs. The BIPM has demonstrated the use of Microsoft's spreadsheet, Excel, in preparing submissions for the MRA Appendix B database, and has circulated a preliminary template to simplify the task of summarizing the results of a Key Comparison in a format suitable for the database [8].

Excel is oriented towards the general spreadsheet user, rather than to the metrologist, although it contains an extensive library of statistical and engineering functions that many scientists routinely use to prepare raw data into graphs and tables. To address special needs, it has some useful customization capabilities. By using the built-in Visual Basic programming environment, it is possible to bring the best of general spreadsheet convenience to the new extensive analyses required for the MRA. We have used this capability to create a toolkit of functions and macros that handle the repetitive tasks associated with evaluating Appendix B data, including particular implementations of confidence intervals for agreement [6].

The toolkit automates aspects of preparation and analysis for Key Comparisons, such as those given below. It can help in preparing MRA Appendix B degree of equivalence tables. It will evaluate confidence intervals ($QDE$) at a wide range of specified confidence, or evaluate the demonstrated confidence ($QDC$) for specified intervals. It organizes and presents them to support the important class of MRA Appendix C decisions that have a directly analogous Key Comparison. In its present form, the toolkit does not offer this level of probability calculus support for less directly supported Appendix C decisions, where careful chains of probabilistic inference from multiple Key Comparisons have to be linked (e.g. length and time for acceleration) and/or combined with intra-laboratory ratiometric uncertainty budgets (e.g. evaluating a 1 kV Appendix C claim using a 10 V Key Comparison and expert opinion about ratiometric practices in the laboratories).

Where calibration and measurement capability listings are supported by comparison data, and for data analysis of Key Comparisons, we believe that the automation provided by this type of toolkit will prove helpful, particularly in coping with the inevitable last-minute additions, changes, and corrections. The toolkit facilitates the exploration of the consequences of different methods for determining the $KCRV$ or its uncertainty. The toolkit can expedite future

review and revision as new Key Comparisons, multilateral regional comparisons, and bilateral comparisons are completed.

Some candidates for the *KCRV*, such as the median and the arithmetic mean are included as built-in functions in Excel, and other statistics used in metrology have been added in the toolkit. One addition is a common choice for evaluating the *KCRV*: the weighted mean, where each laboratory value weight is the inverse of the square of the corresponding standard uncertainty. In addition to this predefined weighted mean function, we have found it useful to have a more general function with "editable weights", that takes as its input a column of values and a column of uncertainties. Weights are initialized to the reciprocals of the squares of the uncertainties, and normalized to unity. It evaluates the inverse-variance-weighted mean, which with normal uncertainty distributions provides the optimal choice, in a minimum-uncertainty sense, for the *KCRV* in a comparison where the values and the uncertainties are all considered to be credible.

This tool also handles the cases where a cutoff is to be used to limit either large uncertainties or large weights for purposes of calculating the *KCRV*. This is done interactively by allowing the user to edit the weight for any laboratory affected by the cutoff. Setting the weight for a given laboratory to zero has the effect of excluding this "outlier" from the evaluation of the *KCRV* without affecting in any way the submissions of that laboratory from other Key Comparison results. Setting a maximum allowable weight can be used to choose a consensus value for the *KCRV* that is not dominated by a single laboratory with a dramatically superior uncertainty claim. The *KCRV* can be chosen to suit the pilot laboratory's convenience, and the traceability of a particular *KCRV* to the SI can be established relative to the independent realizations published in the Key Comparison Report.

With the toolkit an Excel user can easily transform simple tabular data for an *N*-laboratory comparison into its full $N \times N$ bilateral degree of equivalence array, displayed in the BIPM-suggested format [8]. Each element of the array describes the bilateral comparison between a pair of laboratories, giving both the pair difference and the pair difference uncertainty. The full two-dimensional form is convenient to use for addressing bilateral equivalence with the tightest justified range. The full array is necessary especially if there are any significant inter-laboratory correlations to treat, and is usually included in the Appendix B database submission for the Key Comparison. This array approach eliminates the intermediary role of the *KCRV* for interlaboratory equivalence.

Key Comparisons presented in the BIPM database are normally to be given in a simple form (giving comparison results relative to the *KCRV*) as well as the full array of bilateral comparisons (which does not require a *KCRV*). Those who seek simplicity in the comparison of each laboratory with the *KCRV* may have difficulty properly assessing the correlation effects imposed by the fully correlated uncertainty in the *KCRV*. If the *KCRV* uncertainty is included rigorously by adding it in quadrature (less the fully correlated part) to each of the laboratory uncertainties, the resulting tables and graphs are unsuitable for their primary purpose of assuring bilateral equivalence. To achieve this purpose, the user of this "simplified" data is required first to remove the fully correlated uncertainty in the *KCRV*, subtracting it in quadrature. If they really cannot be persuaded to use the full array, we believe that these readers' need for simplicity could be served best by assigning an uncertainty of zero to the *KCRV* (i.e. setting $u_{\mathrm{KCRV}}=0$) when interlaboratory differences are to be reconstructed from the difference with respect to the *KCRV*. (Note that a graph of comparison data shown as deviations from a *KCRV* which has non-zero uncertainty is unsuitable for deducing laboratory-to-laboratory bilateral degrees of equivalence.) We would suggest giving an uncertainty to the SI value of the *KCRV* only in the context of comparisons with the independent realizations and any future reinterpretations of these Comparisons. Nonetheless, the toolkit can handle either choice.

The programming for the macro which generates a full $N \times N$ bilateral equivalence array from a table of comparison data is straightforward for other programmers to use as a template to automate preparation of any other $N \times N$ array. The program itself is documented with this role in mind. We have found it easy to implement calculation of bilateral quantities, such as the one-parameter $QDE_{0.95}$ confidence interval for agreement, or the $QDC(k)$ demonstrated equivalence probability. These applications of confidence intervals and confidence levels are discussed in more detail in the next Section, using data taken from a real comparison. More complex calculations, such as explicitly including any common, correlated uncertainty component (such as a travel uncertainty for the comparison, or the uncertainty due to a prescribed correction formula that may take on different values for different participants), or making use of available correlation coefficient matrix information, are being implemented in the toolkit. We are also adding graphing tools that are useful in presenting Key Comparisons.

Confidence interval calculations can also be made with respect to the *KCRV* (determined in various ways), using the *KCRV* uncertainty (also attributed in various ways) and expressed in an $N \times 1$ table for any particular choice of *KCRV* and $u_{\mathrm{KCRV}}$. We have found it convenient to use the toolkit to treat simultaneously a number, *n*, of candidate *KCRV*s, either as *n* separate $N \times 1$ arrays or by appending them as *n* additional rows (and/or columns) to the $N \times N$ array. The toolkit includes a pair uncertainty module that can accommodate correlation coefficients. It can populate a full two dimensional degree-of-equivalence array, which might have many hundreds of entries, from a Key

Comparison list of a few tens of differences, uncertainties and correlation coefficients. These arrays, along with the associated graphs, are straightforward to understand and use, and the consequences of inter-laboratory correlations are properly reflected in the degree of equivalence entries. Although correlations with the *KCRV* need to be handled on a case-by-case basis when determining the MRA degree of equivalence, the $QDE_{0.95}$ and $QDC(k)$ functions have been automated for use with correlation coefficients.

The utility of automated calculation is best illustrated with examples. The toolkit facilitates the generation of the two-parameter form of the degree of equivalence and its translation into a quantitative confidence for agreement with another laboratory (or agreement with a *KCRV*) within an interval. Given the interval, the confidence can be calculated; and given the target confidence, an interval can be calculated. We have found this to be a powerful tool in guiding deliberations concerning comparisons.

## 6. Examples of Confidence Interval Analysis

Key Comparisons are being conducted under the auspices of the BIPM Consultative Committees in order to determine the degree of equivalence between various NMIs for all of the principal measurement techniques in each of the major metrology areas. Two notions of equivalence are introduced by the MRA. Firstly, the MRA defines the *degree of equivalence* between a laboratory and the *KCRV* as the difference between them, and the expanded ($k$=2) uncertainty of the difference. Secondly, between two laboratories the MRA *bilateral degree of equivalence* is similarly defined to be the difference in their respective comparison values, and the expanded ($k$=2) uncertainty of that difference. Confidence intervals can be calculated in either approach: for all NMIs relative to the *KCRV*, or the *KCRV* can be bypassed and confidence intervals calculated for all bilateral pairs of NMIs.

The first notion, equivalence to the *KCRV*, may be useful in assessing the global consistency demonstrated by a Key Comparison. Tables listing the degree of equivalence to the *KCRV* are included in the MRA Appendix B database that summarizes the results of Key Comparisons. It is sometimes problematic to define a universally accepted *KCRV* for a given comparison, and the toolkit can support a multiplicity of candidate *KCRV*s until the most appropriate one is chosen. Furthermore, even when there is little or no debate about the methodology for determining the *KCRV*, the uncertainty in the *KCRV*, $u_{KCRV}$, often requires careful handling when making further calculations in order to eliminate the effects of correlations with the uncertainties of each of the laboratory values. In practice, the utility of these Key Comparison reference quantities will be limited to the context in which they are intended, as a simple summary of a complicated experiment.

The second notion, the bilateral degrees of equivalence between each pair of participating NMIs, is the equivalence idea of chief practical interest to the participating metrologists, since the vast majority of trade and equivalence issues arefundamentally bilateral. It is also often of paramount concern for those reviewing the calibration and measurement capability claims submitted for inclusion into the MRA Appendix C database, since the bilateral equivalence is the clearest description of the consequences arising from one NMI accepting another NMI's claim. Also, it is the ultimate technical reassurance that metrology can offer for any specific question on equivalence that has not blindly pooled NMIs. The full bilateral array of these degrees of equivalence is also included in the MRA Appendix B database, where the convention is to express (row-*column*) differences, and the uncertainty of that difference, in each array element. Both parts of this array are antisymmetric, since swapping laboratories leads to a simple change of sign for the difference, and $+/-u_p = -(-/+u_p)$ . (Usually the + and – uncertainties are the same and only the magnitude of the uncertainty will be reported so that the uncertainty part will be written to appear to be symmetric.)

We illustrate these ideas by considering in some detail one example typifying a comparison with no major anomalies. The comparison is abstracted from the thermometry literature, where it has not excited commentary as being untypical in any way [10]. For our purposes we may almost ignore the fact that it is a comparison drawn from a particular field: the unit of measurement (the kelvin) has been retained to aid clarity, but the discussion would be the same for most other quantities. The experiment consisted of making temperature measurements of a travelling mercury fixed-point cell, and comparing against the local mercury fixed-point standard. The results were summarized in the natural manner: as a table of the differences ($T_{lab}$ - $T_{circ}$) and the expanded uncertainties (with $k$=2) of the differences. To build the analogy with a Key Comparison, the triple point temperature of the circulating artefact would probably be a candidate for the *KCRV* in this study.

It is often the case that the *KCRV* is calculated as an aggregate statistic determined from the experimental data, rather than being associated with a specific element. For this data set, the median deviation is 0.00 mK, the average deviation is –0.03 mK, and the weighted average (using the experimental uncertainties to determine the weights) is –0.01 mK. The consistency of these values with each other is one indication that there is no significant cause for concern with the data set, and that we should feel free to use any one of them as the *KCRV*. The fact that the

magnitude of these aggregate quantities is so close to zero is indicative of the high quality of the circulating artefact and the consistency among the participating laboratories. Since we believe that the weighted mean will prove to be the most common statistical choice for a reference value in comparisons where the laboratory values and uncertainties are all deemed credible, we will use this as the *KCRV* (adjusting the published table of values accordingly) in the remainder of our examples. With this fictitious assignment of a "*KCRV*", the data shown in the left side of Table 1 can be regarded as typical for an international comparison. According to the MRA, a degree of equivalence table includes the difference between the laboratory value and the *KCRV*, as well as the uncertainty of this difference.

### *6.1 Analysis with $u_{KCRV} = 0$*

In this example we consider the *KCRV* to have been chosen as the weighted mean described above, and take $u_{KCRV} = 0$, which means that the raw laboratory uncertainties are listed directly in the table. For completeness we have listed both the standard uncertainty and the expanded ($k$=2) uncertainty.

If the *KCRV* is viewed from the perspective of a traceable reference to be used for documenting the SI value of each laboratory, an assignment of $u_{KCRV} = 0$ appears unjustifiable. However, it is justifiable within the confines of discussion concerning the equivalence of laboratories to each other, rather than their equivalence to the SI. If the compact form of Table 1 is to be used, taking $u_{KCRV} = 0$ can be a useful simplification for *KCRV*-based presentations, particularly where travel uncertainty is small (and the uncertainty of the *KCRV* is fully correlated for each laboratory). It is also useful when authors wish to avoid difficulties that their readers may experience in accounting for correlations in deducing bilateral comparisons between laboratories from comparisons with the *KCRV*. Further, assigning $u_{KCRV} = 0$ is a technique for communicating a CC's belief that a particular *KCRV* has no lasting physical significance as a record of the SI reference of the comparison.

**Table 1. Comparisons of Laboratories with respect to a *KCRV*, with $u_{KCRV}$=0.**
**Left:** Comparison data summarized by the Degree of Equivalence to the *KCRV*, also showing the standard uncertainty.
**Middle:** Confidence intervals for agreement of the laboratory value *V* and the *KCRV*, at 68% and 95% levels of confidence.
**Right:** *QDC*($k$=2), the demonstrated confidence that a laboratory value includes the *KCRV* in its expanded ($k$=2) uncertainty claim.

| Lab Name | $V_i$-*KCRV* *(mK)* | $u_i(k=1)$ *(mK)* | $U_i(k=2)$ *(mK)* | $QDE_{0.68}$ *(mK)* | $QDE_{0.95}$ *(mK)* | *QDC(k=2)* |
|---|---|---|---|---|---|---|
| Lab1 | 0.01 | 0.13 | 0.25 | *0.13* | *0.26* | *95%* |
| Lab2 | 0.01 | 0.14 | 0.28 | *0.14* | *0.28* | *95%* |
| Lab3 | 0.03 | 0.10 | 0.19 | *0.10* | *0.21* | *94%* |
| Lab4 | -0.09 | 0.08 | 0.16 | *0.13* | *0.22* | *82%* |
| Lab5 | 0.13 | 0.09 | 0.17 | *0.18* | *0.28* | *70%* |
| Lab6 | -0.11 | 0.11 | 0.22 | *0.16* | *0.29* | *85%* |
| Lab7 | 0.18 | 0.13 | 0.25 | *0.25* | *0.40* | *72%* |
| Lab8 | 0.13 | 0.12 | 0.23 | *0.19* | *0.33* | *81%* |
| Lab9 | -0.02 | 0.15 | 0.30 | *0.15* | *0.30* | *95%* |
| Lab10 | -0.08 | 0.16 | 0.32 | *0.18* | *0.35* | *93%* |
| Lab11 | -0.41 | 0.16 | 0.31 | *0.48* | *0.67* | *30%* |

The participants have each measured the same artefact, and reported their measured value and its uncertainty. The results are summarized relative to the *KCRV* in the table. These data are also shown as the solid symbols in Figure 1, with error bars representing both the standard uncertainty $u(k=1)$, and the expanded uncertainty $U(k=2)$. In this analysis, because we have taken $u_{KCRV} = 0$, the usual "visual inspection" of the compatibility of each laboratory result with the *KCRV* would be done by checking to see whether or not the $k$=1 or $k$=2 error bars cross the *KCRV* zero-line. Graphical presentation of the degree of equivalence easily displays both parameters (difference and

uncertainty), and it is tempting to display the *QDE* intervals as well. We have done so in two ways, in Figure 1a and in Figure 1b. These presentations differ only in the positioning of the mid-point of the *QDE* confidence interval: in Figure 1a, the plot is from the perspective of the individual laboratories; in Figure 1b, the plot is from the perspective of the *KCRV*. Both presentations are correct in that these confidence intervals are each centred on the concept of equivalence: which can be centred either on the Laboratory value or on the *KCRV*.

**Figure 1.** Plots of the degree of equivalence with respect to the *KCRV* for the data summarized in Table 1. The $k$=1 and $k$=2 uncertainties are shown as error bars on the solid symbols. The $QDE_{0.68}$ and $QDE_{0.95}$ confidence intervals are shown in the lighter colour as error bars on the open symbols.

**Figure 1a.** Table 1 Comparison Data and *QDE* intervals of the eleven laboratories with each interval centred on its laboratory mean. This shows each laboratory's view of its agreement with the *KCRV*. Note that these *QDE* intervals always cross the *KCRV* line.

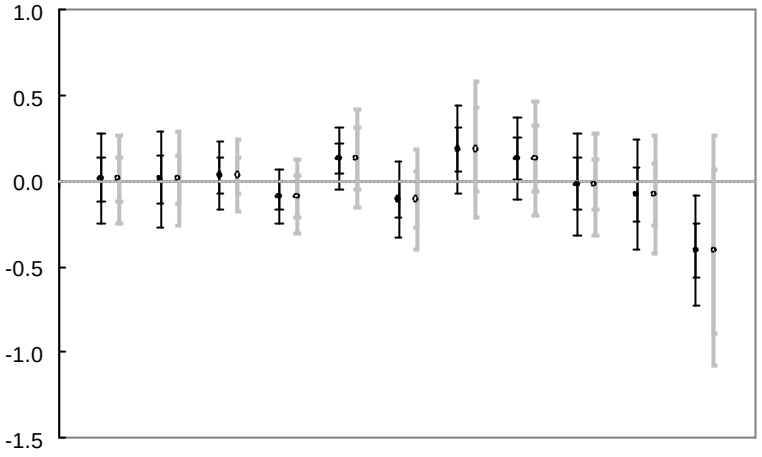
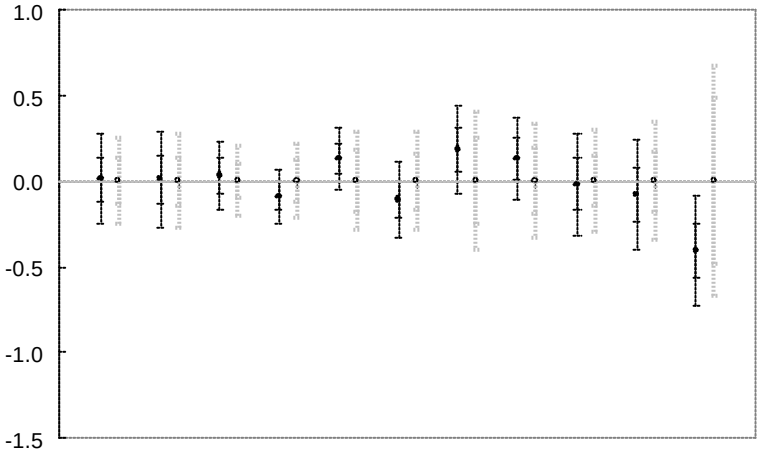



**Figure 1b.** Table 1 Comparison and *QDE* intervals relative to the *KCRV*. The *QDE* intervals show the agreement intervals of the eleven laboratories with the *KCRV*. Note that this is the same information as plotted in Figure 1a, but with the confidence intervals centred on a common international perspective rather than Figure 1a's eleven national perspectives of the *KCRV*. Note also that these *QDE* intervals always encompass the laboratory value.




In choosing to use a graph to display the one-parameter *QDE* confidence interval, it may be necessary for metrologists to help readers troubled by the two different choices. It is important for a reader to remember that a *QDE* interval is not anchored at a particular SI value, but at the concept of agreement between two things. It can never be "wrong" to plot the *QDE* intervals around a particular measurement of either party to the agreement. Figures 1a and 1b are both correct, yet each is subject to misinterpretation if the equivalence intervals are considered simply as error bars. *QDE* intervals are inherently a shared opinion of the two parties, and are always symmetrically

centred on ideal agreement; the graphs in Figure 1 represent the projection of this concept onto the value of one party or the other. This opinion is a shared one, even when the comparison is between a laboratory value and a *KCRV*.

If the unfamiliarity of *QDE* presentations in Figure 1 creates doubts in the minds of any of our clients, we may revert to the form of Table 1: it is in tables and in words that the *QDE* formalism is least subject to misinterpretation. However, comparisons interpreted solely with respect to a *KCRV* are susceptible to a more serious failure of analysis. The greatest weakness of this form of comparison, i.e. using the *KCRV* as an intermediary, is through an explicit or implicit claim of bilateral equivalence between laboratories A and B just because they are individually equivalent to the *KCRV*. Given that A and B individually meet a specific test for equivalence to the *KCRV*, in general one cannot deduce that A and B meet the same test for agreement with each other.

For example, consider the specifics for Lab 4 and Lab 5. Each of these participants agrees with the *KCRV* well within their respective $k$=2 expanded uncertainty claims, and they very nearly agree with the *KCRV* to within their standard $k$=1 uncertainty claim. Their differences with respect to the *KCRV* are similar in magnitude, but have opposite signs: Lab 4 lies below the *KCRV*, while Lab 5 lies above. Based on this type of "graphical" analysis of their performance in the Key Comparison, it is unlikely that many metrologists would question the measurement capabilities or the uncertainty claims of either of these participants: they have done quite well, and both may claim to "agree with the *KCRV*". Yet their $k$=1 uncertainty bars do not touch each other, and a bit more doubt would exist about their equivalence to one another. The bilateral evaluation of *QDC* can quantify the confidence level for agreement demonstrated between participants of the comparison, using the differences and uncertainties.

As discussed in Section 4.2, the *QDE* formalism can be used to evaluate a confidence interval within which each laboratory value will agree with the *KCRV* given the pair uncertainty. The middle portion of Table 1 provides these intervals at confidence levels of 68% and 95%, evaluated using the data in the left portion of Table 1 and Equation 4 with the appropriate coefficients. It is convenient to display these intervals graphically, as has been done in Figure 1a with open symbols. In a manner exactly analogous to showing the $k$=1 and $k$=2 uncertainties, the $QDE_{0.68}$ and $QDE_{0.95}$ confidence intervals are shown there as error bars centred on each laboratory's value. It is worth reflecting on precisely what this form means: relative to each laboratory, the bars represent the confidence interval within which the *KCRV* could be expected to be found if the laboratory repeated its measurement after randomizing all uncertainty components in its uncertainty budget. As such, it is a laboratory-centric view of the comparison and the *KCRV*.

The first thing to note about Figure 1a is that these *QDE* intervals always cross the *KCRV* line. This would be true for all other *QDE* intervals with confidence >50%, calculated for agreement between each laboratory value and the *KCRV*. Secondly, one might note the non-linearity in the underlying (normal) distributions that is evident in the relative size of the 68% and 95% confidence intervals: when the laboratory value does not coincide with the *KCRV*, the 68% confidence interval is noticeably more than half the size of the 95% confidence interval. This is seen most clearly in the data set for Lab 11 in the table and the graph, where $QDE_{0.68}$ is almost three-quarters the size of $QDE_{0.95}$. Contrast this ratio for the *QDE*s of a normal distribution offset from zero, with the case of a normal distribution centred on zero (typically representative of an individual laboratory uncertainty statement, for example) where the standard $k$=1 uncertainty (~68% confidence) is half the size of the expanded $k$=2 uncertainty (~95% confidence).

Returning to the specific examination of Lab 4 and Lab 5, Table 1 may be used to read off their respective $QDE_{0.68}$ and $QDE_{0.95}$ confidence intervals with respect to the *KCRV*: 0.13 mK and 0.22 mK for Lab 4 and 0.17 mK and 0.27 mK for Lab 5. These intervals incorporate both the difference from the *KCRV* and the uncertainty into a single parameter that quantitatively answers the question of how well a laboratory's claimed value agrees with the *KCRV*. For this specific question one must consider both sources of possible disagreement: the difference in value and the uncertainty. For a laboratory whose two sources are comparable, i.e. their value differs from the *KCRV* by about one standard uncertainty, then the 68% confidence interval for agreement is about twice that amount. This is a rigorous consequence of probability calculus centred on the postulated agreement, and a laboratory's conviction that a repeat of its measurement is best described by its stated mean and stated uncertainty.

The *QDE* approach scrupulously preserves the mean and standard uncertainty as separate entities and combines their effects to answer specific questions about equivalence within a specified interval. Most other common approaches do not leave this expert opinion intact, but generally try to supplant it with a pooling assumption; in these approaches the comparison is essentially regarded as a weak one-sample test of an assumed zero deviation within the stated uncertainty (the null hypothesis test). With enough repetition, this null hypothesis type of test can become a very useful approach to evaluating the completeness of the combined standard uncertainty, but from our perspective it has often proved to be a poor approach to interpret infrequently repeated Key Comparisons and equivalence, leading only to weak statements of confidence in equivalence and to unnecessary contention.

As one example of the unnecessary contention, consider the case of a laboratory that has just failed the null hypothesis test at the 95% confidence level, and in consequence is threatened with not being listed as they wish in Appendix C of the MRA. Since up to 1 in 20 laboratories are expected to fall into this category, even if absolutely nothing is wrong, laboratories in this circumstance can be expected to mount an energetic campaign for special treatment.

Proponents of null hypothesis testing could debate what they might regard as double counting in the *QDE* approach, which rigorously separates treatment of the uncertainty and the difference of the means. They might be tempted to regard the measured difference as no more than a single instance of the pair uncertainty. This perspective is founded on the uncritical substitution of the measurement expert's opinion (that a repeated measurement is expected to be distributed around the stated mean value) with an unwarranted pooling hypothesis (that a repeated measurement from this laboratory is expected to be distributed around the stated *KCRV*). As a hypothesis, the pooled interpretation might possibly be tested rigorously with enough repetition. In practice, NMIs do not have the resources to repeat a Key Comparison enough times. The mean would have to be tested to within a small fraction of the stated uncertainty to justify replacing a laboratory's stated mean with the *KCRV*. Null hypothesis testing's need for repetition and the infrequency of Key Comparisons place them at opposite poles of the metrological world.

The last column in Table 1 summarizes the demonstrated confidence, expressed as a percentage, that each laboratory will include the *KCRV* within its claimed $k$=2 uncertainty. It is not surprising that $QDC(k=2)$ = 95% for Lab 1 and Lab 2, which have reported a value virtually equal to the *KCRV*. For Lab 4 and Lab 5, who report differences from the *KCRV* comparable to their $k$=1 standard uncertainty, the demonstrated confidence fractions are 82% and 70%, respectively, which most could feel comfortable accepting. The value of the quantification of confidence is perhaps best illustrated using the results for Lab 11, which has reported a difference from the *KCRV* of almost three standard uncertainties. In this case, $QDC(k=2)$ falls off sharply to only 30%. These *QDC* numerical figures of merit have been calculated with the same rigorous probability calculus as *QDE*, to express the confidence for agreement with the *KCRV* within $\pm 2u$, taking into account the complete information (both the value and the uncertainty) supplied by each participant.

For the purposes of comparing any laboratory to the *KCRV*, summary tables such as Table 1 are sufficient. To summarize the laboratory-to-laboratory bilateral degrees of equivalence, one needs to use the full, antisymmetric, bilateral degree of equivalence table. The CCQM has adopted a form that meets the needs of chemical metrology [8]. An alternative form that is more self-explanatory to physical metrologists is shown in Table 2. In either form, it is a compact summary of all the demonstrated bilateral equivalences for a given experimental comparison. In Table 2, the MRA degree of equivalence is included in the above-diagonal array elements: calculated as the difference between the laboratory values (row – column) and the uncertainty of this difference (the pair uncertainty). Below the diagonal, the 95% confidence interval, $QDE_{0.95}$, for agreement between two laboratories is included (in italics). Since the full MRA-style array is antisymmetric, no information has been lost by including the $QDE_{0.95}$ values in this presentation.

In this illustrative example calculation, the uncertainties in Table 1 have been taken to be uncorrelated when evaluating the laboratory-to-laboratory pair uncertainty. Handling the case of any identified correlations is not difficult, using any specific correlation information provided in the preferred form of interlaboratory correlation coefficients in the Key Comparison report. This presentation of the comparison data makes possible a quick review both of the differences between participants and of the 95% confidence interval within which they are reporting the same value.

**Table 2.** Bilateral Degree of Equivalence and $QDE_{0.95}$ array for the data summarized in Table 1. All entries are in mK.

| | *Lab1* | *Lab2* | *Lab3* | *Lab4* | *Lab5* | *Lab6* | *Lab7* | *Lab8* | *Lab9* | *Lab10* | *Lab11* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lab 1 | - | 0.00 ± 0.38 | -0.02 ± 0.31 | 0.10 ± 0.30 | -0.12 ± 0.30 | 0.12 ± 0.33 | -0.17 ± 0.35 | -0.12 ± 0.34 | 0.03 ± 0.39 | 0.09 ± 0.41 | 0.42 ± 0.40 |
| Lab 2 | *0.37* | - | -0.02 ± 0.34 | 0.10 ± 0.32 | -0.12 ± 0.33 | 0.12 ± 0.36 | -0.17 ± 0.38 | -0.12 ± 0.36 | 0.03 ± 0.41 | 0.09 ± 0.43 | 0.42 ± 0.42 |
| Lab 3 | *0.31* | *0.33* | - | 0.12 ± 0.25 | -0.10 ± 0.25 | 0.14 ± 0.29 | -0.15 ± 0.31 | -0.10 ± 0.30 | 0.05 ± 0.36 | 0.11 ± 0.37 | 0.44 ± 0.36 |
| Lab 4 | *0.35* | *0.37* | *0.33* | - | -0.22 ± 0.23 | 0.02 ± 0.27 | -0.27 ± 0.30 | -0.22 ± 0.28 | -0.07 ± 0.34 | -0.01 ± 0.36 | 0.32 ± 0.35 |
| Lab 5 | *0.37* | *0.39* | *0.31* | *0.41* | - | 0.24 ± 0.28 | -0.05 ± 0.30 | 0.00 ± 0.29 | 0.15 ± 0.34 | 0.21 ± 0.36 | 0.54 ± 0.35 |
| Lab 6 | *0.40* | *0.42* | *0.38* | *0.27* | *0.47* | - | -0.29 ± 0.33 | -0.24 ± 0.32 | -0.09 ± 0.37 | -0.03 ± 0.39 | 0.30 ± 0.38 |
| Lab 7 | *0.46* | *0.48* | *0.41* | *0.51* | *0.31* | *0.56* | - | 0.05 ± 0.34 | 0.20 ± 0.39 | 0.26 ± 0.41 | 0.59 ± 0.40 |
| Lab 8 | *0.40* | *0.42* | *0.35* | *0.45* | *0.28* | *0.50* | *0.35* | - | 0.15 ± 0.38 | 0.21 ± 0.39 | 0.54 ± 0.39 |
| Lab 9 | *0.39* | *0.40* | *0.36* | *0.36* | *0.44* | *0.40* | *0.52* | *0.46* | - | 0.06 ± 0.44 | 0.39 ± 0.43 |
| Lab 10 | *0.44* | *0.45* | *0.42* | *0.35* | *0.51* | *0.38* | *0.59* | *0.53* | *0.44* | - | 0.33 ± 0.45 |
| Lab 11 | *0.75* | *0.76* | *0.74* | *0.61* | *0.83* | *0.61* | *0.92* | *0.86* | *0.74* | *0.70* | - |

Looking at the row for Lab 4 in Table 2, and moving across to the column for Lab 5, the bilateral MRA degree of equivalence, i.e. the difference between their values and the expanded $k$=2 uncertainty of this difference, for this pair of participants is –0.22 ± 0.23 mK, which shows that the two laboratories agree within their mutual expanded uncertainty by only a very small amount. The corresponding $QDE_{0.95}$ entry for this pair is found by locating the column for Lab 4 and moving down to the row for Lab 5: $QDE_{0.95}$(Lab4,Lab5) = 0.41 mK . Note that it is always possible to calculate the $QDE_{0.95}$ interval (or any $QDE_{C\%}$ interval) for agreement at the 95% confidence level, even when the values do not "overlap" within the error bars.

In order to demonstrate the utility of $QDC(k=2)$, the complete bilateral demonstrated confidence array has been tabulated in Table 3. Note that it is not symmetric, since the laboratory uncertainty claims are different from each other. The full $QDC(k=2)$ matrix quantifies the confidence which one should construe that **Lab (row)** and ***Lab (column)*** will agree within **Lab (row)**'s $k$=2 claim.

**Table 3.** Table 1 Bilateral Confidence array for $QDC(k=2)$ . Each entry gives the confidence level for reproducing agreement between **Lab i** (row) and ***Lab j*** (*column*) within the **Lab i** (row) $k$=2 expanded uncertainty claim.

| | *Lab1* | *Lab2* | *Lab3* | *Lab4* | *Lab5* | *Lab6* | *Lab7* | *Lab8* | *Lab9* | *Lab10* | *Lab11* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lab1 | - | 83% | 88% | 84% | 80% | 78% | 68% | 77% | 80% | 75% | 22% |
| Lab2 | 86% | - | 89% | 86% | 82% | 80% | 71% | 79% | 82% | 77% | 25% |
| Lab3 | 77% | 75% | - | 73% | 76% | 65% | 60% | 71% | 71% | 63% | 10% |
| Lab4 | 61% | 59% | 61% | - | 31% | 76% | 23% | 33% | 61% | 63% | 18% |
| Lab5 | 62% | 61% | 71% | 37% | - | 33% | 72% | 77% | 54% | 42% | 2% |
| Lab6 | 70% | 68% | 70% | 89% | 44% | - | 34% | 45% | 71% | 74% | 34% |
| Lab7 | 68% | 67% | 74% | 47% | 88% | 43% | - | 84% | 61% | 49% | 5% |
| Lab8 | 73% | 72% | 80% | 55% | 89% | 50% | 81% | - | 66% | 55% | 7% |
| Lab9 | 86% | 85% | 89% | 90% | 80% | 85% | 69% | 77% | - | 81% | 34% |
| Lab10 | 84% | 83% | 86% | 93% | 72% | 90% | 61% | 70% | 84% | - | 48% |
| Lab11 | 31% | 32% | 26% | 50% | 12% | 54% | 10% | 14% | 37% | 48% | - |

Once again it is interesting to consider the specific case involving Lab 4 and Lab 5. Above it was shown that they each agree quite well with the *KCRV*. Here, the demonstrated confidence that each of these participants includes the other's value within their respective expanded uncertainty claims (and taking explicit account of the pair uncertainty) has been evaluated to be only about 35%. This reflects the fact that the two values, although both only about one standard uncertainty away from the *KCRV*, are on opposite sides of this baseline. This was also apparent in their bilateral MRA degree of equivalence, which had a difference in value equal to the ($k$=2) uncertainty of that difference. In terms of the "normalized error", the ratio of the difference to the expanded uncertainty which is often used in null hypothesis testing, this pair of laboratories would have scored $E_n$(Lab 4, Lab 5) = 0.96. This score, although close to the nominal cutoff for acceptability, would normally lead to a statement indicating that Lab 4 and

Lab 5 are "equivalent". In contrast, the use of $QDC(k=2)$ has shown that there is only a small probability that these two laboratories agree within their respective uncertainty claims, and serves as a strong warning about the quality of judgements which may be made when using more traditional acceptance criteria.

### *6.2 Subtleties of Analysis with $u_{KCRV} \neq 0$*

We turn next to the case where the *KCRV* has a non-zero uncertainty assigned to it and where this has been accepted by the responsible Consultative Committee. In some Key Comparisons, $u_{KCRV}$ will be evaluated as the combined standard uncertainty of the data used to construct the *KCRV* itself. Such a choice for $u_{KCRV}$ would describe the expected reproducibility of the *KCRV* under specified assumptions regarding correlations between laboratories and between repetitions. In other comparisons, $u_{KCRV}$ might be taken as the standard deviation of the comparison data to describe the breadth of the overall pooled distribution. In still other comparisons, $u_{KCRV}$ might be dominated by a travel uncertainty, which is evaluated during the comparison itself. Where the *KCRV* is chosen to be the value of the pilot laboratory, $u_{KCRV}$ may logically be taken as the pilot laboratory uncertainty. For each of these cases, a detailed description of $u_{KCRV}$ is required for the complete report on the Key Comparison, and the Consultative Committee might well also document the appropriate uses of the particular $u_{KCRV}$.

One of the subtleties with a non-zero $u_{KCRV}$ is the issue of correlations between the uncertainty components used by the different laboratories. The subtleties do not necessarily complicate things. For example, in any situation where $u_{KCRV}$ is deemed to be completely correlated with all of the participants, the bilateral analysis of equivalence is identical to the case where $u_{KCRV}$ is taken to be identically zero. This uncertainty cancels out during the evaluation of the pair uncertainty due to the fact that a difference between two values is being calculated. The lab-to-lab quantities such as the bilateral degree of equivalence, $QDE_{0.95}$ and $QDC(k=2)$, remain invariant. This would be the case for a comparison where the uncertainty in the *KCRV* is dominated by a single instance of an uncertain quantity. As an example, consider the uncertainty in the Josephson constant relative to the SI definition of the volt, which would be the same for all laboratories – i.e. fully correlated in uncertainty, with correlation coefficients of +1 for this uncertainty component. In describing the uncertainty in this way, it is important that the readers are all made aware of the change of sign required when using the correlation coefficient in a difference compared to using it in a sum: $u_p^2 = u_1^2 + u_2^2 - 2\, r_{12} u_1 u_2$ for a difference $V_2 - V_1$, and $u_p^2 = u_1^2 + u_2^2 + 2\, r_{12} u_1 u_2$ for a sum $V_1 + V_2$.

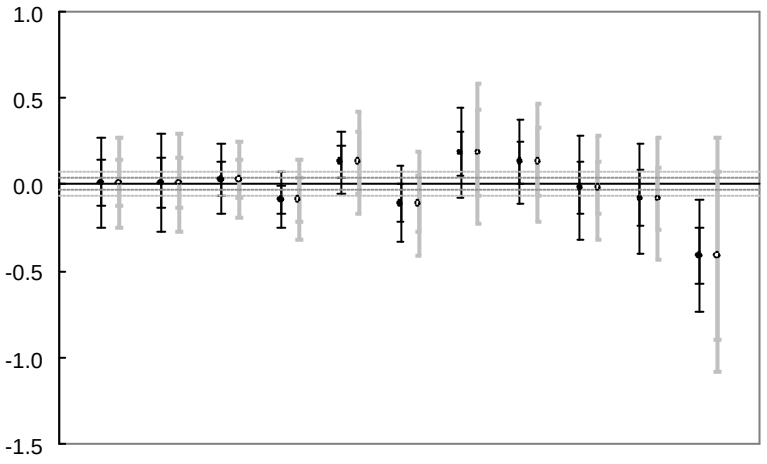
### *6.3 Analysis with $u_{KCRV}$ as the Formal Uncertainty of the Inverse-variance Weighted Mean*

To illustrate calculations with an explicit uncertainty assigned to the *KCRV*, we use the same comparison data listed in the left-hand part of Table 1, but take $u_{KCRV} = 0.03$ mK. This value is equal to the combined standard uncertainty of the weighted mean as calculated above, and is obtained by taking the quadrature sum of the laboratory uncertainties with the appropriate weighting factors. Table 4 lists the confidence intervals and demonstrated confidence for agreement with this new *KCRV*, exactly as was done in the middle and right-hand portions of Table 1. Figure 2 provides the same graphical summary of the data as Figure 1, but now includes dotted lines to indicate $u_{KCRV}$. The calculation of the pair uncertainty used for the error bars in the figure has also changed to include the new value for $u_{KCRV}$, but the laboratory uncertainty claims as listed in Table 1 remain unaltered.

**Table 4.** Confidence for the comparison of Table 1, but with $u_{KCRV}$ = 0.03 mK. $QDE_{0.68}$ and $QDE_{0.95}$ are the intervals within which the laboratory value and the *KCRV* agree, at 68% and 95% levels of confidence. $QDC(k=2)$ represents the demonstrated confidence that the laboratory values include the redetermined, "smeared out", *KCRV* within their expanded ($k$=2) uncertainty claims.

| Lab Name | $QDE_{0.68}$ (mK) | $QDE_{0.95}$ (mK) | $QDC(k=2)$ |
|---|---|---|---|
| Lab 1 | *0.13* | *0.26* | *95%* |
| Lab 2 | *0.14* | *0.28* | *95%* |
| Lab 3 | *0.11* | *0.22* | *93%* |
| Lab 4 | *0.13* | *0.23* | *80%* |
| Lab5 | *0.18* | *0.29* | *68%* |
| Lab 6 | *0.17* | *0.30* | *84%* |
| Lab 7 | *0.25* | *0.41* | *71%* |
| Lab 8 | *0.20* | *0.34* | *80%* |
| Lab 9 | *0.15* | *0.30* | *95%* |
| Lab 10 | *0.180* | *0.35* | *92%* |
| Lab 11 | *0.48* | *0.68* | *30%* |

**Figure 2.** Table 1 equivalence with respect to a *KCRV* with $u_{KCRV}$=0.03 mK (indicated by the dotted lines). The data points are the same as in Figure 1a, but the error bars now include the small but non-zero uncertainty of the *KCRV*. The error bars represent the combined laboratory and *KCRV* uncertainties, with coverage factors of $k$=1 and $k$=2 (dark lines, solid symbols). Note that this presentation of the data is unsuitable for making judgements about the bilateral degree of equivalence. The $QDE_{0.68}$ and $QDE_{0.95}$ confidence intervals for agreement with the *KCRV* (light lines, open symbols) are also shown, and take account of the uncertainty in the *KCRV*.




This $u_{\mathrm{KCRV}}$ is substantially smaller than the uncertainty of any one laboratory, and would describe the expected variation in the *KCRV* if the Key Comparison were to be repeated with all influence parameters fully randomized. If the participants have identified the reproducibility part of the uncertainty, the CC could reasonably choose to use this even smaller uncertainty for the KCRV, and for the laboratory uncertainties. Explicitly accounting for the unavoidable correlations could further reduce $u_p$, the pair uncertainty between any one laboratory and the *KCRV* [7].

### *6.4 Analysis with $u_{KCRV}$ as the Standard Deviation of the Participants from the KCRV*

The Consultative Committee might instead decide to use $u_{\mathrm{KCRV}}$ to describe the width of the pooled distribution of laboratory results, determined perhaps from the external standard deviation. With our example data set, the standard deviation of the data points is 0.16 mK with 10 degrees of freedom (taking the Laboratories' results to be independent). Such an approach might appear attractive if the repeatability-of-the-*KCRV* approach discussed in the previous section is felt to be misleadingly narrow, and has a rigorously justifiable interpretation that will be described below. Table 5 lists the confidence intervals and demonstrated confidence for agreement using this $u_{\mathrm{KCRV}}$, as was done in Table 1, for infinite degrees of freedom.

We have also used the Welch-Satterthwaite estimate of the degrees of freedom for the pair uncertainty distribution (Lab *i* versus the pool). For the eleven participants it ranges from 14 to 41. When the explicit degrees of freedom are considered [3], the $QDE_{0.95}$ estimates of Table 5 are increased by less than 9%. We can also use the stated standard uncertainties in forming the pool estimate. When the pooled distribution (the sum of the eleven normal distributions, each centred on its own mean as revealed in the Key Comparison) is explicitly created, convoluted and integrated, the $QDE_{0.95}$ estimates of Table 5 increase by less than 25%.

Figure 3 shows the data of Table 5 in the same graphical summary as Figures 1a and 2. The calculation of the pair uncertainty used for the error bars in the figure includes the new value $u_{\mathrm{KCRV}}$=0.16 mK, but the laboratory uncertainty claims as listed in Table 1 remain unaltered.

There is a rigorous interpretation attached to this approach. The pooled distribution is not intended to describe the reproducibility of the *KCRV* for a fully randomized repetition, but rather for a similar and fully complete repetition of the Key Comparison. Nor does this $u_{\mathrm{KCRV}}$ attempt to estimate the uncertainty with respect to the SI. In that case, one would have to include the uncertainty of the thermodynamic temperature of the triple point of mercury among other things. Instead, this wider distribution is attempting to describe the reproducibility of a measurement drawn randomly from the pool of Laboratories that participated in the Key Comparison. The imagined repetition is a bilateral comparison between the laboratory in question, and an anonymous, randomly chosen, second participant.

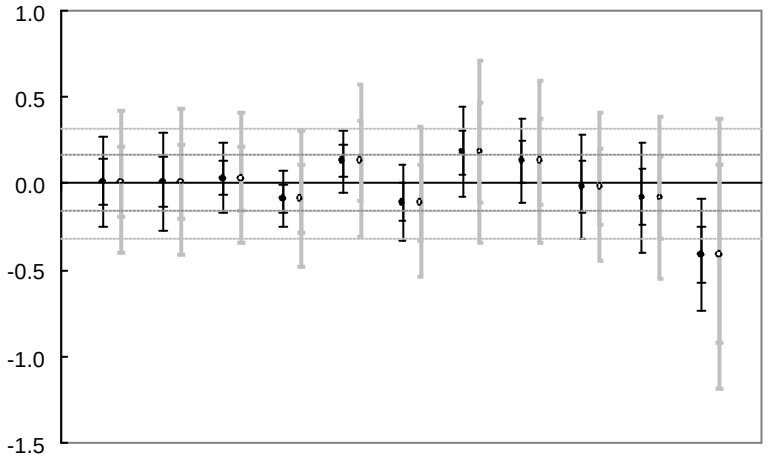
With this non-negligible $u_{\mathrm{KCRV}}$, all of the confidence intervals, at both the 68% and 95% levels of confidence, have increased in comparison to the values in Tables 1 and 2. The increase is the direct result of increasing $u_{\mathrm{p}}$, which now includes a non-zero contribution from $u_{\mathrm{KCRV}}$. The $QDC(k=2)$ values for all of the participants except Lab 11 have decreased, again due to the increase in the pair uncertainty. Lab 11, which differs from the *KCRV* by almost three of its own standard uncertainties, has benefited from the "smearing out" of the *KCRV*, and the demonstrated confidence has increased from 30% to 35%. Unfortunately the increased confidence is still quite low, and a price has been paid by all of the other participants. The highest cost has been paid by the laboratories which agreed most closely, and had the smallest difference in value from the *KCRV*: the demonstrated confidence that might be claimed for Lab 1 has dropped from 95% to 79% as a result of increasing $u_{\mathrm{KCRV}}$ from 0.00 mK to 0.16 mK.

The increased uncertainty weakens confidence most for laboratories demonstrating "better" overlap with the *KCRV*, and does little to help the "worst" laboratory. The lab-to-lab results, included in the bilateral arrays of Section 6.1, remain unaffected, under the assumption that $u_{KCRV}$ is fully correlated. We believe that the *KCRV* should have little or no general relevance outside of the scope of the Key Comparison report. In this context, the example shows the quantitative effect of not attributing an uncertainty to the reference value. It quantifies how much, and which, NMIs lose and gain. This example argues against the suggestion that $u_{KCRV}$ might be artificially inflated to "rescue" Laboratories with outliers in the comparison by this subtle variant of guardbanding.

**Table 5.** Confidence for the comparison of Table 1, but with $u_{KCRV}$ = 0.16 mK. $QDE_{0.68}$ and $QDE_{0.95}$ are the intervals within which the laboratory value and the *KCRV* agree, at 68% and 95% levels of confidence. $QDC(k=2)$ represents the demonstrated confidence that for a similar measurement, the laboratory values include, within their expanded ($k$=2) uncertainty claims, the value from a randomly chosen contributor to the *KCRV*.

| Lab Name | $QDE_{0.68}$ (mK) | $QDE_{0.95}$ (mK) | $QDC(k=2)$ |
|---|---|---|---|
| Lab 1 | *0.21* | *0.41* | *79%* |
| Lab 2 | *0.21* | *0.42* | *81%* |
| Lab 3 | *0.19* | *0.38* | *70%* |
| Lab 4 | *0.20* | *0.39* | *57%* |
| Lab5 | *0.24* | *0.44* | *55%* |
| Lab 6 | *0.22* | *0.43* | *67%* |
| Lab 7 | *0.29* | *0.53* | *63%* |
| Lab 8 | *0.25* | *0.47* | *67%* |
| Lab 9 | *0.22* | *0.43* | *83%* |
| Lab 10 | *0.23* | *0.47* | *82%* |
| Lab 11 | *0.51* | *0.78* | *35%* |

**Figure SEQ ARABE3.** Table 1 equivalence with respect to a *KCRV* with $u_{KCRV}$=0.16 mK (indicated by the dotted lines). The data points are the same as in Figure 1a and Figure 2, but the error bars now include the larger uncertainty of a randomly chosen contributor the *KCRV*. The error bars represent the combined laboratory and *KCRV* uncertainties, with coverage factors of $k$=1 and $k$=2 (dark lines, solid symbols). Note that this presentation of the data is unsuitable for making judgements about the bilateral degree of equivalence. The $QDE_{0.68}$ and $QDE_{0.95}$ confidence intervals for agreement with the *KCRV* (light lines, open symbols) are also shown, and take account of the uncertainty in the *KCRV*.




INCORPORER

## 7. Confidence Interval Analysis for Appendix C Submissions

For a more typical Appendix C evaluation, the appropriate calibration service uncertainty claims from each laboratory (which may be different from their Key Comparison uncertainty given in Appendix B) would be used to calculate the confidence matrix summary of the Appendix B Key Comparison. Note that each laboratory is most interested in its horizontal row (here, others' acceptance of its own Appendix C claim) and its vertical column (its acceptance of others' claims).

In order to construct a defensible acceptance criterion for use when evaluating Appendix C calibration and measurement capabilities, one might consider using this application of the *QDE* formalism and calculate the probability that a laboratory's measurement includes the *KCRV* within the expanded uncertainty associated with its Appendix C claim.

In this scenario, it would be up to the individual NMIs, within each of the Regional Metrology Organizations (RMOs), to determine their "cutoff" level of confidence, below which Appendix C claims will be returned to the submitting Institute for re-analysis and review. A criterion for acceptance can be stated clearly, evaluated rigorously, and applied uniformly. An example of such a statement follows.

> *At (our NMI), for services directly related to Appendix B Key Comparisons, we accept an Appendix C claimed $U_{CAL}$ provided the Key Comparison demonstrated confidence within $\pm U_{CAL}$ for agreement with the KCRV is at least x%.*
>
> *i.e. QDC(k=2,Lab,KCRV) > x%*

Each NMI can instead use its own realization rather than the *KCRV* as its arbitrating reference. This is a policy decision with technical, political and legal overtones, and the use of one's own standard simplifies most of these issues. Acceptance criteria may well be the subject of debate with the RMO or Consultative Committee. For all the constituent bilateral acceptance agreements, the most defensible choice is an NMI's own standard, making no reference at all to an external reference quantity (the *KCRV*) which may have been calculated using input data from several other laboratories who are not party to the particular bilateral agreement being considered. In this case, a statement of the acceptance criterion might read as follows.

> *At (our NMI), for services directly related to Appendix B Key Comparisons, we accept an Appendix C claimed $U_{CAL}$ provided the Key Comparison demonstrated confidence within $\pm U_{CAL}$ for agreement with our value as reported in the Key Comparison is at least x%.*
>
> *i.e. QDC(k=2,Lab, our NMI) > x%*

Once an NMI or RMO has chosen a value ($C=x\%$) for its minimum demonstrated confidence, the toolkit makes it easy to evaluate Appendix C claims if there is a corresponding Key Comparison in Appendix B. This subset of Appendix C claims is important. The technical foundations of the MRA require that the principal techniques in each area of metrology be fully investigated experimentally. It is therefore vitally important that the claims for this special subset of cases should be reviewed and analyzed completely and rigorously. In this context, the toolkit simplifies the use of confidence levels (*QDC*) and confidence intervals (*QDE*) to examine the consistency and values of the choices for *x*% made by other NMI's and RMO's. With confidence quantifies in this way, decisions can be explained clearly to users for whom the MRA databases are being created.

### *7.1 Quantifying Confidence Assigned to Expert Opinion*

It is difficult to quantify the exact role of Key Comparisons in the quality management system for international metrology, and yet the expense of Key Comparisons will have to be justified in the budgets of many NMIs. As shown below, when a Key Comparison is available, it is sometimes possible to quantify the extent to which reliance is placed on supplementary information.

A policy that relies solely on the confidence developed by the Key Comparison would ignore many sources of confidence: other information, other comparisons, history, judgement and expert opinion. A consistent policy for Appendix C acceptances would use the same level of confidence in all decisions in a field. An NMI might make some acceptances at levels of demonstrated confidence lower than for some rejections. Unless it is being deliberately inconsistent, that NMI is exhibiting the reliance that it is placing on other information. Appendix C acceptances provide a large pool of decisions that might be examined to determine quantitatively the role of supplementary information when there is an exactly matching Appendix B listing.

The role of confidence derived from supplementary sources can sometimes be inferred from the spread of demonstrated confidence in the extensive set of decisions that an NMI has made. Suppose a particular NMI accepts Laboratory A's claim with a Key Comparison *QDC* of *a*%, and rejects Laboratory R's claim with a *QDC* of *r*%; then if $r>a$, it is clear that this NMI must be assigning a role to the supplementary information with at least a confidence of $(r-a)$%.

To illustrate this quantification of the role of expert opinion with the above example, suppose that the MRA Appendix C entries being considered were all at the highest level: i.e. the entries have the same uncertainties as the

results summarized in the Key Comparison above. This implies that the confidence matrix for these particular Appendix C decisions is that of Table 3. (Usually, the Appendix C claims will be at a lower level - having substantially higher levels of confidence demonstrated by the Key Comparison - but this does not affect the illustrative power of this example.) Recall that in Table 3 the column for each Laboratory tabulates, for each of the other row Laboratories, the confidence demonstrated by the Key Comparison that repetitions of the comparison would fall within the row Laboratory's $k$=2 claim. Thus each Laboratory needs to consider only its own column when voting on their acceptance of the other Laboratories' claims, and needs to consider only its own row when arguing for the acceptance of its own claim by the other Laboratories. Suppose that Laboratory 6 has accepted Laboratory 5 (33% confidence from Table 3) and not Laboratory 4 (76% confidence from Table 3): these Laboratory 6 decisions reveal that it is allowing its additional information or its expert opinion to contribute at least (76-33)=43% confidence. The difference may be due to additional confidence that could come from additional bilateral comparisons, confidence from Laboratory visits and judgement on Laboratory procedures, different confidence calculated from its own correlation analysis (and different than that used for Table 3), or even knowledge of breakdowns in the Key Comparison. Whatever the origins for this additional confidence that Laboratory 6 has in the capabilities of Laboratory 5, the *QDC* method can quantify the minimum role played by supplementary information and expert opinion in acceptances for Appendix C of the MRA: *"expert opinion has contributed at least 43% confidence to decisions made by Laboratory 6"*. This process can be applied to a single NMI in a single Key Comparison, to the pooled NMIs in a Key Comparison, to a single NMI in a pool of related Key Comparisons, to a single NMI in a pool of unrelated Key Comparisons, or to decisions made or ratified by a Regional Metrology Organization. Key Comparisons are expensive, and quantifying the balance between their role and the role of other information will allow for more informed decisions regarding participation in Key Comparisons in support of inclusion in Appendix C.

If a Laboratory has identified a particular confidence level that must be reached to accept an Appendix C submission, then any deficits revealed by their column (and row) of *QDC* values may be remedied by increasing the Appendix C interval, or by relying on other information sources and expert opinion to cover the deficit. This treatment of all sources of information is explicit in the MRA itself, but the use of the *QDC* method has allowed us to treat the decisions quantitatively.

## 8. Discussion

Metrologists are used to dealing with extensive collections of measurements from within their own laboratories, where they are often making decisions about what data is to be pooled and what data is to remain unpooled. Typically, only one metrologist's opinion about the measurements needs to be described. Interlaboratory comparisons are fundamentally different, since measurements from more than one laboratory are being considered. The bilateral difference and pair uncertainty incorporate opinions held by at least two different metrologists.

Even with their intra-laboratory experience, metrologists are finding significant challenges in interpreting the wealth of interlaboratory comparisons being compiled for Appendix B in support of decisions on inclusion in Appendix C. Intra-laboratory intuitions concerning data pooling are not easily transportable to the inter-laboratory context. The consequences of excluding one data point can be very different in the two contexts. Within a single laboratory, the exclusion of a data point from a large set of otherwise indistinguishable points would normally be expected to be benign from every perspective in metrology. In the inter-laboratory case, however, it could have the effect of excluding an NMI's calibration and measurement capabilities from Appendix C. The rigor required to make these decisions for Appendix C is of a different, higher order than normally required in intra-laboratory analyses, since the impact of dropping "one point" is so much higher.

The appropriate starting point for examining equivalence rigorously is to treat the data as being independent, rather than prematurely pooling the comparison data. $QDE_{0.95}$ and $QDC(k=2)$ allow the demonstrated level of equivalence to be calculated rigorously, and provide a solid basis for decisions during the review process: either as "*Yes, include this claim - No, exclude that claim...*" or as "*For approval into Appendix C, only claims ±U¢ or greater will be accepted"*.

The other element that is facilitated in the confidencelevel or confidenceinterval approach is the natural attachment of ownership to opinions expressed by combinations of probability distributions. In a measurement made by a particular NMI, the probability distribution is centred on the mean and is of a width normally expressed by the standard uncertainty; it is the opinion of that NMI's responsible metrologist. In a bilateral comparison, expressed as a difference of two measurements made in two NMI's, the two responsible metrologists must examine and express their own uncertainties, and may be able to examine each other's uncertainty budgets for correlations. The opinion expressed in the two-parameter degree of equivalence form is then the joint opinion of the two responsible

metrologists. In a multilateral comparison, the pilot laboratory may be the best source for estimating the travel uncertainty to attach to each bilateral pair of comparisons in the multilateral comparison, and in this case the joint opinion regarding each bilateral comparison will be a three-party opinion. When a *KCRV* is chosen, a part of the opinion expressed will come from those choosing the weights for the different measurements, and part will come from those responsible for each NMI's measurement, and some might come from the pilot laboratory expert in assigning travel or local measurement uncertainty. Comparisons of an NMI's results with the *KCRV* are an expression of the most diffuse source of opinion, while for bilateral comparisons it will often be easier to identify whose opinion is being expressed. Correlations will normally be quite high between the uncertainty of the difference between one NMI's value and the *KCRV*, and the uncertainty of the difference for another NMI. These correlated components will largely cancel out when bilateral comparisons are calculated, even when using a *KCRV* intermediary.

There is another opinion that is worth mentioning in this context, and that is the uncertainty distribution created in the mind of an attentive reader. Communicating expert opinion to the reader to assist him in forming his own opinion is the sole purpose of all the work done in performing, analyzing, and reporting Key Comparisons. Impediments in communicating details to the reader are to be avoided, and means to facilitate understanding and belief need development.

When the most difficult questions arise in deliberations surrounding the MRA, it may prove helpful to identify ownership of the opinions expressed by probability distributions and their combinations. The confidence interval and confidence level approaches make this clearer. An NMI facing rejection of its calibration and measurement capability claims from Appendix C can properly look for support in the bilateral comparisons with other laboratories. Each of the other laboratories will reveal the basis for its own opinion about that NMI in its bilateral comparison abstracted from the Key Comparison. These joint opinions on the expectation of agreement are expressed most clearly by the confidence levels for agreement within the Appendix C claim. From an *N*-laboratory multilateral comparison, each laboratory will have *N*-1 joint opinions for agreement with the *N*-1 other NMIs. An NMI facing rejection has the option of attempting to obtain higher confidence from the other NMIs by broadening its Appendix C claim for agreement, and this points out another strong need for a rapid means of recalculating these aspects of probability calculus.

Metrologists today are facing the first wave of calibration and measurement capability submissions arriving for evaluation both within and among the six Regional Metrology Organizations. Some metrologists are only beginning to appreciate the intricacies of this new task. For example, when $u_{\mathrm{KCRV}}$ is non-zero, equivalence acceptance cannot rely on a simple interpretation of equivalence to the *KCRV*: it is not transitive (i.e., where ≡ denotes equivalence, $A \equiv K$ and $B \equiv K$ does not imply that $A \equiv B$). In this situation, the best strategy is to compile bilateral equivalence tables, which summarize all of the laboratory-to-laboratory degrees of equivalence.

The task of examining complex comparison experiments that are summarized in Appendix B of the MRA as tables of differences, with uncertainties, is very large, and will grow larger. A justifiable evaluation criterion is a first step towards uniform application of a technically supported peer review process. Any organization which can articulate a confidence level which is deemed sufficient for acceptance will be able to use $QDC(k=2)$ to facilitate acceptance decisions.

Equation 3 and the $QDC(k=2)$ notation are specific to normal distributions. For non-normal distributions, a new version of Equation 3 must be integrated, and a new notation for *QDC* would be needed that would suggest that the claimed capability $\pm U$ is being made at, say, a claimed confidence level of 95% (denoted by the $k=2$ in our notation for normal distributions). The demonstrated confidence level for agreement with the arbitrating reference within $\pm U$, is *x%* computed by the new Equation 3, and perhaps can be much less than 95%. Distinguishing these two percentages (the claimed confidence and the demonstrated confidence) is a notational challenge when both are needed. However, the sentence describing the result remains simple: "*Within the interval having a claimed confidence of 95%, the confidence level demonstrated by the comparison is x%.*"

Other criteria, even those that appear to be less well specified (such as a graphical analysis based on error bars crossing the *KCRV*) can be translated using the *QDE* formalism into an explicit confidence level. Thus a decision can be scrutinized with full mathematical rigor and applied to any other case with *ISO Guide*-compliant uncertainty statements in any major metrology area, including degrees of freedom, correlations, and non-normal distributions.

## 9. Conclusion

We have illustrated the use of confidence interval methods in describing careful comparison measurements and their associated uncertainty budgets. We have shown confidence intervals to be particularly useful in describing CIPM Key Comparison data, and in supporting decisions made concerning submissions made for Appendix C of the MRA.

We have shown how these techniques can even be used to quantify the minimum role assigned to expert opinion, when expert opinion is blended with Key Comparison data for Appendix C decisions. A toolkit was introduced for using confidence intervals with these MRA submissions, and has been shown to simplify the daunting tasks in describing the essentials of the data. We conclude that Key Comparison data, representing the acme of the metrologists' art, warrant the most careful statistical treatment. Rigorous confidence interval techniques are exactly fitted to the demands of MRA decision-making. Even for the extensive decisions of Appendix C, confidence intervals can be used easily with an appropriate toolkit.

**References**


1. CIPM, Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, (1999).
2. ISO Guide to the Expression of Uncertainty in Measurement, (1992).
3. Wood B.M. and Douglas R.J., Confidence-interval interpretation of a measurement pair for quantifying equivalence, *Metrologia* **35** (1998) pp. 187–196.
4. Wood B.M and Douglas R.J., Quantifying Demonstrated Equivalence, *IEEE Trans. Instrum. Meas.*, **48** (1999) pp. 162-165.
5. Steele A.G., Wood B.M. and Douglas R.J., Quantifying equivalence for interlaboratory comparisons of fixed points, Proceedings of *Tempmeko'99*, (1999) pp. 245-250.
6. Steele A.G., Wood B.M. and Douglas R.J., Quantifying Demonstrated Equivalence: A *QDE* Toolkit, in *Advanced Mathematical and Computational Tools in Metrology* **V**, World Scientific, Lisbon Portugal (2001).
7. Steele A.G., Wood B.M. and Douglas R.J., Exclusive statistics: simple treatment of the unavoidable correlations from KCRVs, to appear in *Metrologia* **38**(6), 2001.
8. Thomas C., Excel Template for Appendix B Submissions, BIPM (2000).
9. Steele A.G., Wood B.M. and Douglas R.J., Interpreting Key Comparison Data for the MRA Database, in *Advanced Mathematical and Computational Tools in Metrology* **V**, World Scientific, Lisbon Portugal (2000).
10. Hermier Y., Bonnier G., Chimenti V., Perezagua F., del Campo D., Weckström T., Marcarino P., Dematteis R., Filipe E., Lobo I., Rauta C., Dubbeldam J.F., De Groot M., Head D.I., Rusby R., Wittwer A., Münch W., Edler F., Fellmuth B., Ivarsson J., A Euromet comparison of the mercury fixed point, reported in the Euromet nº280 Synthesis Report (1997).


**Appendix**

Recall the "master equation" used in probability calculus for evaluating confidence levels and confidence intervals: the confidence level $C$, in per-cent, for agreement within an interval $\pm d_C$, symmetrically located around zero (i.e. the expectation of agreement) is obtained by integrating the pair uncertainty distribution for the difference over the interval.

$$C\% = \int_{-d_C}^{+d_C} P_p(z)dz \cdot 100\% \qquad \text{(A.1)}$$

For a given level of confidence $C$, this integral equation can be solved for the half-interval $d_C$ by iteration. For a comparison between two normal distributions, with means $m_1$ and $m_2$, and a pair standard uncertainty $u_p$, there is an approximation formula that may be used to facilitate calculations of confidence intervals.

$$QDE_{C\%} \approx \; |m_1 - m_2| + \{a_0(C) + a_1(C) \times \exp[-a_2(C) \times (|m_1 - m_2|)/u_p]\} u_p \qquad \text{(A.2)}$$

Coefficients for use with the approximation formula are given in Table A.1. They have been adjusted to fit iterative numerical solutions of Equation A.1. The Maximum Error column gives the maximum difference between each parameterized fit and the full numerical solution. At high confidence levels, the maximum error introduced by equation A.2 is small relative to the usual limitations arising from the lack of certainty regarding the distribution and its tails. With equation A.2, the best fit of the interval's breadth at lower confidence levels has larger errors, particularly where $|m_1 - m_2| << u_p$. We do not believe that there is wide utility in calculating the intervals at these low confidence levels, but if ever they are needed, equation A.1 can be solved at any required level of accuracy, again limited by the lack of certainty regarding the distribution.

**Table A.1**. The coefficients for the *QDE* approximation, Equation A.2; listed for different confidence levels, *C*. The column labelled *"Maximum error of Equation A.2"* is the maximum value of | *QDE* approximation - *QDE* interval | / (*QDE* interval), for any combination of $z=|m_1-m_2|/u_p$ and $u_p$. For all confidence levels, the maximum error tends to zero as $z$ becomes greater than about $2u_p$. For lower confidence levels ($C<0.5$), the error in the Equation A.2 approximation is large, but only for $z$ near 0.

| *C* | $a_0(C)$ | $a_1(C)$ | $a_2(C)$ | *Maximum Error of Equation A.2* |
|---|---|---|---|---|
| 0.995 | 2.576 | 0.236 | 5.287 | *0.2%* |
| 0.95 | 1.645 | 0.3295 | 4.050 | *0.5%* |
| 0.90 | 1.282 | 0.375 | 3.595 | *0.8%* |
| 0.85 | 1.036 | 0.418 | 3.300 | *1%* |
| 0.80 | 0.842 | 0.458 | 3.076 | *1%* |
| 0.75 | 0.674 | 0.498 | 2.890 | *2%* |
| 0.70 | 0.524 | 0.537 | 2.728 | *2%* |
| 0.68 | 0.468 | 0.554 | 2.669 | *3%* |
| 0.65 | 0.385 | 0.578 | 2.584 | *3%* |
| 0.60 | 0.253 | 0.622 | 2.450 | *4%* |
| 0.55 | 0.126 | 0.668 | 2.326 | *5%* |
| 0.50 | 0.000 | 0.718 | 2.207 | *6%* |
| 0.45 | -0.126 | 0.773 | 2.092 | *8%* |
| 0.40 | -0.253 | 0.834 | 1.980 | *11%* |