Confidence-interval interpretation of a measurement pair for quantifying a comparison

B. M. Wood and R. J. Douglas

Abstract. We present a method for applying the results from a pair of comparison measurements, made in two laboratories, to form a confidence interval needed to support equivalence statements. Only the usual assumptions indicated in the ISO *Guide to the Expression of Uncertainty in Measurement* about the interpretation of the means and combined standard uncertainties are required. The method is mathematically rigorous and can be used to calculate the interval to arbitrary precision for any confidence level. It can also include the effects of correlations and degrees of freedom. Simple graphical and numerical methods of practical accuracy are presented for commonly used confidence levels. The method permits statements for clients to be made of the form "On the basis of comparison measurements [reference] performed in the period of [date to date], the results of similar measurements made at [Laboratory 1] and [Laboratory 2] can be expected to agree to within $(\pm d_{0.95})$, with 95 % confidence". We derive and discuss the first rigorous justification for this type of statement.

1. Introduction

The *Guide to the Expression of Uncertainty in Measurement* [1] forms the basis for communicating measurement uncertainties. It recommends the use of the combined standard uncertainty for expressing the uncertainty in a measurement. However, neither the *Guide*, nor currently published practices of national [2, 3] or international metrology organizations [4-6], recommend a single-parameter description to express the degree of agreement between the measurement results of even the simplest comparison. The simplest comparison consists of just a pair of measurements of an exchanged artefact, one measurement made in each of two laboratories. We refer to this as a measurement pair and develop a way to evaluate and express its information.

The full information of any set of measurements is contained in the original data and the detailed uncertainty budget prepared by the metrologist involved. The great usefulness of the *Guide* is to structure an uncertainty budget such that all the uncertainty components can be combined in a consistent and reasonably rigorous manner to describe the mean and the combined standard uncertainty of the set of measurements. This is a two-parameter representation of the total information in a set of measurements. Sometimes, these two parameters must be supplemented with the effective degrees of freedom and with covariances. This multi-parameter description of a set of measurements is the most commonly accepted form in metrology.

B. M. Wood and R. J. Douglas: National Research Council of Canada, Ottawa, K1A 0R6 Ontario, Canada.

Despite the wide acceptance of the *Guide*'s multi-parameter description, other descriptions of the total information are also useful, especially for presentation to a more general audience which does not want to be concerned with all (or any) of the metrological complexities. The *Guide* recognizes this fact, and recommends the use of a confidence interval to describe a measurement and its uncertainty as the confidence that the measurement is contained within the quoted interval. The *Guide* goes further and recommends the terminology of a coverage factor, k, to calculate an appropriate confidence interval.

Metrological comparisons are performed to test and quantify the agreement of measurements performed in different laboratories. Typically, an artefact is transported in turn to several laboratories, to have its particular quantity measured by each laboratory. The measurements are referred to the same system of units, usually the International System of Units (SI), although different techniques of realization may be used in the different laboratories. The comparison may be useful in determining whether any difference is due to mistakes, differences in interpretation, random occurrences or unknown causes; but the primary purpose of the comparison is still to test and quantify the agreement of measurements. While it may be hoped or even believed that the laboratories are measuring the same quantity in the same units, the greatest utility of the comparison can be obtained if we allow for the possibility that the measurements may be very different. To allow for the possibility of a difference, we do not assume that measurements from different laboratories are sampled from a single distribution. Each laboratory is sampling from its own distribution and we initially treat each laboratory's distribution as independent no matter how

similar it may be to another laboratory's distribution. Departures from independence are treated rigorously by considering correlations in the uncertainty budgets.

A multi-laboratory metrological comparison can be interpreted in many ways. For example, an interval which encompasses all the results might satisfy some of the less-demanding users, but it does not express how closely the more tightly grouped laboratories might agree. Creating subsets of this interval, sometimes referred to as banding, could be done, but again these intervals do not properly express how closely a smaller set or pair of laboratories might agree. The only way to describe the smallest interval of agreement warranted by the particular measurements is to consider each measurement pair. In the most general case this includes all the pairs of measurement results of the comparison. Of course, this requires up to $N \times (N-1)/2$ descriptions for a comparison involving N laboratories. Another type of comparison pair could consist of an individual laboratory's measurements compared with an arbitrary value, such as an interlaboratory mean, determined with various weightings for the different laboratories. Each weighting would lead to another N comparison pairs to be considered, each a minor variation of the same pair-comparison problem. Comparisons among twenty or more laboratories are not uncommon and this emphasizes the need for an easily calculated, rigorous and widely accepted quantification of the equivalence of the measurement pairs. The effort put into these comparisons and their importance demand that the interpretation of each measurement pair be defined as well as the measurement results warrant.

Any interpretation of comparison measurements is constrained by what actions either laboratory takes on the basis of the measurement results. In particular, laboratories may, or may not, decide to adjust their reference values on the basis of the results of the comparison. These two cases are different and must be interpreted differently. This paper considers a method applicable to the general situation in which each laboratory does not alter its reference values after the comparison. The case in which one laboratory adjusts its reference value by the difference between its results and another laboratory's results can be treated as a particular instance of the general method.

Our goal is to reach the broadest clientele, for whom it is most appropriate to use a confidence interval describing the agreement of a pair of measurement results. We propose the following principles for the interpretation of comparison results. The interpretation should:

- quantify the level of agreement;
- describe the best possible agreement warranted by the measurement results;
- be able to treat each measurement result independently: each having its own distribution;
- accommodate any measurement pair described in accordance with the *Guide*, even those which fail the usual null-hypothesis tests [4-6];
- be able to treat effective degrees of freedom, covariance terms and non-normal distributions;
- be expressed as a single parameter to communicate most effectively the level of agreement.

We adopt the basic perspective of the *Guide*, that the measurements of the comparison are each described by a parameterized distribution. It is important to consider the sensitivity of this interpretation with respect to these parameters. Although no general interpretation can recreate the information lost by mistakes that perturb the means or by unrealistic or incorrect estimations of the uncertainties, our approach is robust in some cases of concern. The consequences of invalid assumptions of the means and the uncertainties are reviewed in the discussion of Section 3.

2. Confidence interval analysis of a measurement pair

Confidence intervals provide a rigorous and easily understandable single-parameter description of a comparison using concepts developed in the *Guide*. Confidence intervals are not recommended by the *Guide* for expressing the uncertainty in a single measurement, but are reserved for use as a tool to interpret the uncertainty for the broadest possible client base. In the context of measurement comparisons, the confidence-interval interpretation avoids the conceptual difficulties of interpreting the comparison as a pass/fail test of the null hypothesis.

Consider the measurement results of just two participants of a comparison. We refer to this as a measurement pair. For this comparison suppose that each laboratory has measured the same artefact, and the measurements result in two means and two uncertainty budgets. Each budget details the combined standard uncertainty, the effective degrees of freedom and sufficient information so that any important covariant items common to each uncertainty budget can be identified. There should also be details of any probability distributions that are not assumed to be normal.

Each laboratory's parameters are used to define a continuous probability distribution best describing the measurement. Typically each distribution will be normal, centred on the mean, m_i, with the combined standard uncertainty, u_i, characterizing the width of the distribution and having effective degrees of freedom v_i. Non-normal distributions may be treated as discussed in Appendix A, if they are adequately described as recommended in the *Guide.* Covariances are also treated in Appendix A where the equations are derived more rigorously and include covariance components. The effect of degrees of freedom is described separately in

Appendix B. Both effects can be determined as well as they are characterized within the individual uncertainty budgets but in the interests of clarity we have omitted these effects from this section.

The continuous probability distributions describing the measurements are $P_1(x)$ for Laboratory 1 and $P_2(y)$ for Laboratory 2. $P_1(x)$ and $P_2(y)$ are allowed to be independent and thus use different arguments, x and y. The probability that Laboratory 1 measures the quantity x between x and $x + \mathrm{d}x$ is $P_1(x)\,\mathrm{d}x$ and similarly for Laboratory 2. We are interested in the probability distribution of the difference of a pair of measurements, estimated from the measurement pair with $x - y$ as the variable. If $z = x - y$ then the probability that z lies between z and $z + \mathrm{d}z$ is $P_p(z)\,\mathrm{d}z$. In the absence of correlations P_p is a convolution of P_1 and P_2. If P_1 and P_2 are uncorrelated normal distributions, then P_p is also a normal distribution centred at $z = m_1 - m_2$ and with a width characterized by the root-sum-square of the combined standard uncertainties, $u_p = \sqrt{(u_1^2 + u_2^2)}$. Appendix A re-derives this probability distribution more rigorously and illustrates it graphically.

The probability distribution, P_p, is now adequately characterized and ready for interpretation. As indicated in the introduction, we wish to determine the confidence interval within which the means of Laboratory 1 and Laboratory 2 would agree. Such an interpretation could, for example, give assurance that the probability that $m_1 = m_2$ within $\pm d_C$ is 95 %. The confidence interval is determined by integrating the probability distribution symmetrically from 0 to $\pm d_C$ until the desired probability, C, is reached. It is important to note that the integration is not centred on the peak of the distribution so the confidence interval, d_C, is dependent on both the difference between the means and the combined standard uncertainties (see Figures 1 and 2).

For the usual type of comparison, the degrees of freedom, ν_1 and ν_2, will be large. In this case the original estimates about the detailed shape of P_1 and P_2 will be unchanged by considerations of the sampling limitations. If, as is generally the case, P_1 and P_2 were assumed to be normal distributions then they can still be represented by the same normal distributions. Consideration of the effects of smaller degrees of freedom is given in Appendix B. For normal distributions, at a particular degree of confidence C, the confidence-interval is $\pm d_C$, and is determined by solving the equation

$$C = \int_{-d_C}^{+d_C} P_p(z)\,\mathrm{d}z = (1/2)\,(\mathrm{erf}\,\{[d_C - (\,|\,m_2 - m_1\,|\,)]/(u_p\sqrt{2})\} + \mathrm{erf}\,\{[d_C + (\,|\,m_2 - m_1\,|\,)]/(u_p\sqrt{2})\}). \qquad (1)$$

The variation of the confidence interval, d_C, with normalized difference of the means, $|\,m_2 - m_1\,|/u_p$, is shown in Figure 1, for three confidence levels: 68 %, 95 % and 99.5 %. The confidence interval starts at the expected values near one sigma, two sigma, and three sigma, and approaches asymptotes of $|\,m_2 - m_1\,| + 0.468\,u_p$, $|\,m_2 - m_1\,| + 1.645\,u_p$, and $|\,m_2 - m_1\,| + 2.576\,u_p$.

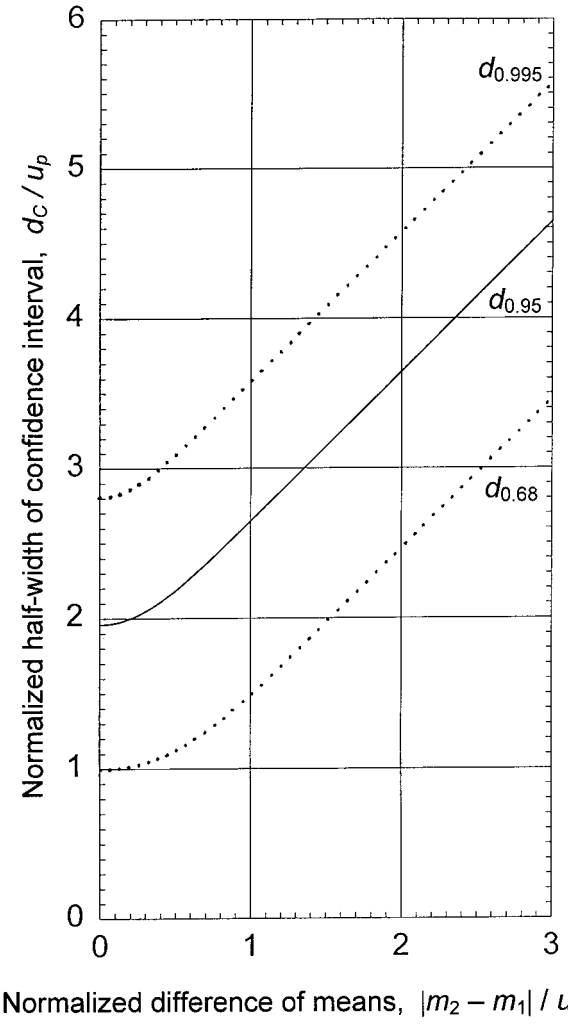

Figure 1. Quantified equivalence for 68 %, 95 % and 99.5 % confidence. A measurement pair with means m_1 and m_2; and with a combined standard uncertainty $u_p = (u_1^2 + u_2^2)^{1/2}$ gives the confidence interval, $(-d_C, +d_C)$. The variation of d_C/u_p with normalized difference $|\,m_2 - m_1\,|/u_p$ is plotted for the three confidence levels.

We believe that the 95 % confidence level is appropriate for most clients' needs and for that confidence we have fitted a function which is simple to calculate and is accurate to better than 1 % of $d_{0.95}$:

$$d_{0.95} \approx |\,m_2 - m_1\,| + \{1.645 + 0.3295 \times \exp[-4.05\,(\,|\,m_2 - m_1\,|\,)/u_p]\}\,u_p. \qquad (2)$$

Equation (2) allows the confidence interval to be simply determined with a hand calculator or spreadsheet

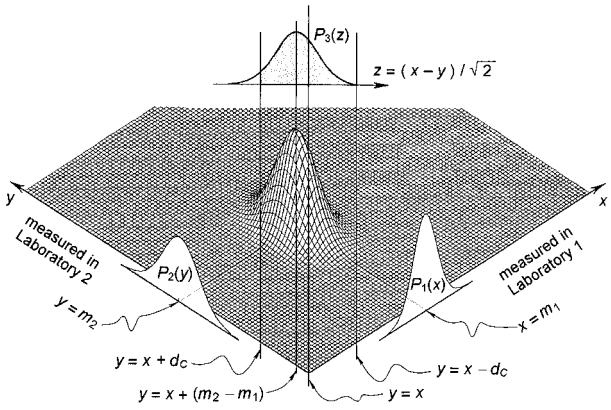

Figure 2. Probability density $P(x, y)$ for a pair of measurements x and y, with means m_1 and m_2. $P(x, y)$ is the complete description of the joint probability. Following the *Guide*, $P(x, y)$ may be estimated from the two means, the standard uncertainties and their correlation coefficient. The confidence level for quantified equivalence within a band $\pm d_C$ is obtained by integration of the probability density in this band, centred on $x = y$. Also shown are the projections of $P(x, y)$ onto the probability densities for measurements by the first laboratory $P_1(x)$, and by the second laboratory $P_2(y)$; and for the measurement difference $P_3(u)$ (scaled to fit this diagram).

program. Figure 1 or (2) allow the confidence interval to be determined simply with sufficient accuracy for any practical comparison.

As an illustration, it may be useful to consider the method as it applies to two extreme situations.

(a) If $m_1 = m_2$ then the confidence interval is determined by integrating a normal distribution, which has a width characterized by $u_p = \sqrt{(u_1^2 + u_2^2)}$, symmetrically about the peak of the distribution. As one would expect, the confidence interval, d_C, is given by the coverage factor times the root-sum-square of the combined uncertainties.

(b) If $|m_2 - m_1| \gg u_p$ then the expected difference of the means is $m_2 - m_1$ and is independent of the confidence level that is sought. Thus $d_C \rightarrow |m_2 - m_1|$ if the difference between the means is very large, again as one would normally expect.

There is one caution concerning the use of any confidence interval: *the confidence interval must not be treated as if it were a standard uncertainty*. As shown in Figure 1, the scaling of d_C from one confidence level to another depends on u_p, on $|m_2 - m_1|$ and on both the initial and final confidence levels. It cannot simply be scaled with a multiplicative factor. For clients requiring an understandable assurance of equivalence, most would be satisfied with the choice of $C = 0.95$, this being the commonest confidence level used in the literature and promoted by the *Guide*.

Figure 1 also illustrates that the normalized difference of the means: $|m_2 - m_1| / u_p$ (or "normalized error", which is used as a figure of merit for null-hypothesis tests [4-6]) is not suitable as the single parameter for quantifying equivalence. The normalized error can be used with a criterion for equivalence, but still requires u_p to interpret the equivalence in units of the measurand. For doing this, a simple "rule of thumb" has been proposed [7]. It approximates two limiting cases: $|m_2 - m_1| / u_p \ll 1$ and $|m_2 - m_1| / u_p \gg 1$. A "degree of equivalence" is assigned the value $2\,u_p$ for comparisons which have $|m_2 - m_1| / u_p < 1$, and is assigned the value $|m_2 - m_1|$ for comparisons which have $|m_2 - m_1| / u_p > 1$. Our analysis shows that this rule of thumb estimates the 95 % confidence interval correctly for the two asymptotic cases, but also can significantly underestimate it. The worst case is at $|m_2 - m_1| = 2\,u_p$ where it is 0.54 of the true value. In return for the small computational burden of (2), [or (B2) to account for finite degrees of freedom], our method also offers the advantages of:

- rigour for making explicit statements to clients about demonstrated confidence for equivalence;
- universality for describing all measurement comparisons for which the uncertainties in both measurements are adequately described in accordance with the *Guide*; and
- efficiency in promptly and usefully interpreting all results, even those which fail null-hypothesis tests.

2.1 Application – interpretation of international comparisons by pair analysis

International comparisons are regularly used to assess and monitor measurement equivalence in different countries. This is of importance in trade, manufacturing and other sectors relying on high accuracy in any measured parameter. National standards laboratories are usually involved at the highest levels of measurement accuracy and international comparisons, typically organized by the Bureau International des Poids et Mesures through the appropriate consultative committee, are performed for important quantities to demonstrate that such measurements are in agreement. The national laboratories have developed different levels of capabilities to service the needs of their countries and as a consequence comparison results often vary widely in their uncertainty claims and occasionally in their mean values. The interpretation of international comparisons using this confidence-interval method has the following advantages.

- The method provides a rigorous confidence interval describing the agreement of the means between every pair of participants.
- It can use simple uncertainty budgets, and yet can accommodate fully detailed uncertainty budgets and rigorously treat correlations, finite degrees of freedom and non-normal distributions.

- It can be adapted to different levels of confidence as required.
- It allows equivalence between two parties to be quantified as closely as the measurement pair results warrant.
- It does not impose on any pair an inflated agreement interval that includes other participants' larger uncertainties.
- It applies equally well to laboratories with very similar means and uncertainties, and to laboratories with very different means or uncertainty claims.
- It does not require knowledge of the "key comparison reference value" or details of how this value was derived.
- The procedure is simple, very easily calculated and applicable to any type of comparison.
- The procedure does not require committee evaluation, but instead directly uses the information as provided by the metrologists involved, i.e. the participants in the comparison.
- The procedure provides a quantified equivalence even for comparisons which would fail to give any formal agreement with interpretations [4-6] based on the null hypothesis.
- The procedure easily re-evaluates comparisons at a later date for audits, for new staff and for new agreements.
- The procedure can be easily adapted to include other parties at a later date (see below). This will be very important as comparisons become larger and include different regional metrology groups.

The appropriate consultative committee may choose to adopt some kind of interlaboratory mean value with which to reference all the comparison results. This value may be defined as the mean of the comparison results, either as a weighted or as an unweighted mean. Constraints to avoid excessive reliance on the results of any one laboratory may also be used. This type of reference value can have a real significance as a consensus value of the SI unit, or instead it might be regarded as an arbitrary, interim value when consensus has not been reached. This value would normally have its uncertainty and its effective degrees of freedom specified. Unfortunately, it will also have correlations with each of the constituent measurement results. In a typical multilaboratory comparison the consultative committee would usually have to consider all of these parameters. Our measurement pair analysis can easily determine a confidence interval in this situation by considering each laboratory in turn, compared with the reference value as the other "laboratory". One consequence of adopting a reference value is to increase the total number of measurement pairs from $N(N-1)/2$ to $N(N+1)/2$. For some applications a lesser number of measurement pairs might suffice. The N comparisons of each laboratory with the mean value, taken as a set, may superficially appear inviting; but in single-parameter form they are not capable of quantifying the best possible agreement warranted by the measurements of any two participating laboratories, nor can they treat significant covariances correctly.

2.2 Application – assessment of two laboratories indirectly through a third common laboratory

This method may be applied to quantifying the equivalence of two Laboratories (1 and 2) indirectly, by each performing comparisons with a common third or "vertex" Laboratory (3). Thus direct comparisons are not required for quantifying equivalence between N laboratories in a study, and $(N-1)$ bilateral comparisons can generate $N(N-1)/2$ bilateral equivalence statements. Furthermore, if a new laboratory wishes at a later date to demonstrate equivalence with these N laboratories, the N new equivalence statements may be derived from just one new measurement pair.

The mean value m_3 of the vertex laboratory will have uncertainty u_3. In the comparison with Laboratories 1 and 2, using a pair uncertainty of $\sqrt{(u_1^2+u_2^2+2\,u_3^2)}$ will usually overestimate the uncertainty. The uncertainty contribution of Laboratory 3 may be better than this due to its stability. Its stability, represented by a smaller standard uncertainty u_{3S}, is estimated from its uncertainty budget by removing (in quadrature) all uncertainty terms known to be fully correlated between the first and second measurements in Laboratory 3, but retaining the travel uncertainty, all randomly sampled uncertainties and any re-sampled distribution which is not known to be fully correlated. Inversely correlated terms will be rare, but must not be removed since they will contribute $4\,u_i^2$ to the variance of the difference. Partially correlated uncertainty terms can be treated by separating them into correlated and uncorrelated parts.

Let $P_1(x) \otimes P_2(y)$ be the distribution obtained by the convolution of $P_1(x)$ and $P_2(y)$. The stability of a measurement in Laboratory 3 is described by the distribution $P_{3S}(z)$ with a standard uncertainty of u_{3S}. The difference in two measurements in Laboratory 3 will be described by the distribution $P_{3S}(z) \otimes P_{3S}(z')$, after all known deterministic biases have been removed. It will have a standard uncertainty $(\sqrt{2})\,u_{3S}$. Then the distribution expressing the comparison of Laboratories 1 and 2 through Laboratory 3 is given by $(P_1(x) \otimes P_{3S}(z)) \otimes (P_2(y) \otimes P_{3S}(z')$. The confidence interval for the equivalence of Laboratories 1 and 2 is obtained by symmetrical integration of this distribution about $x = y$. For normal distributions, the combined standard uncertainty for comparisons through a vertex laboratory is $u_p = \sqrt{(u_1^2+u_2^2+2\,u_{3S}^2}$. To calculate a confidence interval from u_p and $|m_2 - m_1|$ (really

$|(m_2 - m_3') - (m_1 - m_3)|$ and assuming $m_3 = m_3'$) we may use Figure 1, or (1), or (2).

Indirect comparison allows a group of laboratories to quantify equivalence efficiently. Indirect comparison gives a somewhat larger d_C than might have been obtained in a direct comparison. If the vertex laboratory has stability equal to or better than the entire uncertainty budget of the other two laboratories, then the indirect d_C can be up to $\sqrt{2}$ worse than for the direct method. This will be the case if travel uncertainty dominates. If the stability of the vertex laboratory (including travel) is half the overall uncertainty of the other two laboratories, then the degradation is no worse than 12 %, or 5 % if it is predominantly travel uncertainty. Normally the vertex laboratory and the artefact will be chosen to minimize this degradation. A somewhat larger d_C may be considered a small compromise to reduce the comparison measurement workload by a factor of between $N/2$ and N (which is often more than 10).

3. Discussion

It is interesting to note that this one-parameter description of equivalence will not have the character of a metric in a vector space, and so could not rigorously support an ordering of comparisons from "best" to "worst". It only describes equivalence and thus can convey no information concerning whether, for example, it is "better" to arrive at a given level of equivalence with a small difference in the means and a large uncertainty, or with a larger difference in the means and a smaller uncertainty. This type of question does not impinge on equivalence, and remains a subject for investigation by metrologists, who would probably be happier with the former case and regard the latter case as implying that there is something in at least one uncertainty budget which is not being evaluated properly.

Because the effective degrees of freedom is never infinite, the consequence of ignoring the effect of a limited number of degrees of freedom is to underestimate the size of the combined uncertainty used to calculate the confidence interval. A rough estimate of that underestimation is obtained from the ratio of the integrals of the Student's t distribution $t_{0.95}(\nu)/t_{0.95}(\infty)$ (see Table G.2 of the *Guide*). The correct treatment is presented in Appendix B.

Because almost all correlations have a positive value of correlation coefficient and represent a relatively small fraction of the total uncertainty, the typical consequence of ignoring the effect of correlations is to obtain a slightly larger confidence interval d_C than would have been obtained if the correlations were properly assessed. For fully correlated individual terms, the correlation can easily be handled in the distribution of the difference by removing fully correlated uncertainty terms from the two uncertainty budgets. A detailed treatment is presented in Appendix A.

It is important to consider the implications of interpreting the confidence interval of a bilateral comparison with contentious results. When contentious results are reported from two metrologists – each of whom has claimed to have given his best estimate of the mean, standard uncertainty and degrees of freedom – a third party can still properly construe d_C as the joint best estimate of the confidence interval for agreement. Since each metrologist has claimed that his results are "correct", or at least his best estimate, then the only additional assumption that the first metrologist must make to quantify the equivalence is that the second metrologist is reporting correctly his mean, standard uncertainty and degrees of freedom.

In comparisons, it is usually possible to employ unknown artefacts to ensure that the mean is reported without bias, and so on average the difference of the means will be properly accounted for within the confidence interval. There is less assurance with respect to standard uncertainties, where unreconciled disagreements are not unknown. For describing equivalence between the two laboratories, the greatest concern of the first metrologist is probably that the second metrologist might underestimate or under-report his standard uncertainty. Interestingly, in the case where two laboratories have roughly equal capabilities, the confidence interval is relatively robust to possible under-reporting of the standard uncertainty by one laboratory. If one laboratory underestimates its uncertainties, then on average d_C should reflect this fact by having included the variation in the difference in the means, as well as the other laboratory's reported uncertainty. For two otherwise identical comparison results the effect of underestimation of one of the uncertainties would result in, at most, a reduction of $1/\sqrt{2}$ (about 30 %) of the confidence interval. We believe that a 30 % relative precision is usually sufficient for the practical purposes of equivalence agreements. This robust characteristic can be a significant consideration in minimizing concerns about the degree of equivalence demonstrated by a comparison.

4. Conclusions

For any measurement comparison between two laboratories, a single parameter can be determined that quantifies their bilateral equivalence in a form suitable for equivalence agreements. The confidence interval can rigorously incorporate both the standard uncertainty and the difference in the means revealed by comparison measurements, as well as the effects of correlations and of limited degrees of freedom. This single-parameter interpretation is also applicable to comparisons utilizing an interlaboratory mean.

The measurement pair analysis naturally leads to an interpretation for multilaboratory comparisons without the need for an interlaboratory mean value. Each

member of a group of N participating laboratories determines $N-1$ individually quantified equivalence statements. Each of these bilateral statements is robust in the sense that it is unaffected by the uncertainty of a third party. Each of the $N-1$ statements is in a form suitable for approval as an equivalence agreement. By working through a vertex laboratory, a participating laboratory may obtain $N-1$ equivalence statements after making only one measurement.

For the pilot laboratory making N or perhaps fewer measurements, the preparation of the $N(N-1)/2$ confidence intervals can be wholly automated and yet incorporate the consequences of known interlaboratory correlation and the effects of limited degrees of freedom.

Acknowledgements. It is a pleasure to thank our colleagues at the National Research Council of Canada for their encouragement and for many helpful discussions on equivalence quantification. In particular we wish to thank G. Chapman, J. Decker, K. Hill, D. Inglis and A. Steele. Opinions in this paper are those of the authors and should not be assumed to imply official policy of the NRC.

Appendix A

Derivation of the confidence interval with covariances included

In this Appendix we derive (1) in sufficient detail to make its rigour accessible to anyone familiar with elementary probability and calculus. We believe that this level of detail is warranted by the wide variety of technical backgrounds represented by clients who are served by equivalence statements, and their possible scepticism of results which they cannot simply find in statistics textbooks.

The statistical basis underlying the *Guide* is that for measurement of a physical variable x, the sources of measurement uncertainty are to be evaluated by the measurer and quantitatively represented by a continuous probability density function, $P_1(x)$. The subscript 1 is used to denote measurements at Laboratory 1. Usually the appropriate functional form is a normal or Gaussian distribution $P_1(x) = \exp[-(x-m_1)^2/2u_1^2]/\sqrt{(2\pi)}\,u_1$, where m_1 is the mean of the measurements of x and where u_1 is the standard uncertainty in x.

For comparisons, we also consider measurements made by Laboratory 2 on the same artefact. These measurements are allowed to be independent of $P_1(x)$ and measure what is construed to be the same physical quantity, but referred to by the variable y and subscript 2. The measurement uncertainty is evaluated and is represented by the continuous probability density function $P_2(y)$. Usually a normal distribution is appropriate here, too: $P_2(y) = \exp[-(y-m_2)^2/2u_2^2]/\sqrt{(2\pi)}\,u_2$, where m_2 is the mean of the measurements of y and u_2 is the standard uncertainty in y. It is easy to see that the standard uncertainties u_1 and u_2 are different in general, so that P_1 and P_2 must be different distributions. It is more subtle to appreciate that for quantifying equivalence, one should also initially allow m_1 to be different from m_2, and that P_1 and P_2 must be different distributions in this respect. The distributions are regarded as distinct in at least these two respects.

For all possible pairs of measurements (x,y), the two-measurement generalization of the *Guide*'s probability density is a continuous function of both x and y, $P(x, y)$. The probability of obtaining a measurement between x and $x+\mathrm{d}x$ is $P_1(x)\,\mathrm{d}x$, and the probability of obtaining a measurement between y and $y+\mathrm{d}y$ is $P_2(y)\,\mathrm{d}y$. In the absence of correlations between x and y, for obtaining both a measurement from Laboratory 1 between x and $x+\mathrm{d}x$ and a measurement from Laboratory 2 between y and $y+\mathrm{d}y$, the probability is the product of the probability $P_1(x)\,\mathrm{d}x$ and the probability $P_2(y)\,\mathrm{d}y$. Thus, in the absence of correlations, $P(x, y) = P_1(x)\,P_2(y)$. The relationship between $P_1(x)$, $P_2(y)$ and $P(x, y)$ is illustrated in Figure 2. $P_1(x)$ is just the integral of $P(x, y)$ over all ways of generating the same x (i.e. the integral over all y); and similarly $P_2(y)$ is just the integral of $P(x, y)$ over all ways of generating the same y (i.e. the integral over all x).

Even in the presence of correlations, $P_1(x) = \int P(x, y)\,\mathrm{d}y$ and $P_2(y) = \int P(x, y)\,\mathrm{d}x$. The effect of correlation between the measurements of two laboratories may be evaluated by simply examining pairs of uncertainty components u_{1j} and u_{2j}, one from each of the two uncertainty budgets. For each component pair the effect of correlation is considered as it affects $P(x, y)$, in particular its effect on u_p. Each pair will have a correlation coefficient r_{ij} and

$$u_p = \sqrt{\left[\Sigma_{i,j}\left(u_{1i}^2 + u_{2j}^2 - 2\,r_{ij}\,u_{1i}\,u_{2j}\right)\right]}. \qquad \text{(A1)}$$

Most pairs will be considered to be uncorrelated, and for these $r_{ij} = 0$. In the context of interlaboratory comparisons, the use of statistical methods to establish correlation coefficients will usually be impractical. For some pairs, however, simple inspection of the two uncertainty budgets will reveal a good estimate of the correlation coefficient.

For example, if both uncertainty budgets have included the same component representing the same travel uncertainty of the artefact then the effect of this component pair has been "double counted" with respect to u_p. This effect of correlation could be treated by removing the component from one budget or by means of the correlation coefficient of this pair of components $r_{ij} = 1/\sqrt{2}$.

Consider another example involving a comparison of two optical frequency standards using a stable travelling laser. Both uncertainty budgets could

include the same uncertainty component owing to the uncertainty of the absolute SI frequency of the travelling laser. However, the uncertainty of the absolute SI frequency of the travelling laser has no effect on the uncertainty of the difference. For this pair of components $r_{ij} = 0$. This could also be treated by considering modified uncertainty budgets for this measurement pair in which this component is removed from the budgets of both laboratories.

It is often convenient to characterize the correlation coefficients, r_{ij}, and then to use a global r with u_1 and u_2 in the calculation of u_p:

$$u_p = \sqrt{[(u_1^2 + u_2^2 - 2\,r u_1\,u_2)]},$$

with

$$r = [\Sigma_{i,j}(r_{ij}\,u_{1i}\,u_{2j})]/\sqrt{[(\Sigma_i\,u_{1i}^2)\,(\Sigma_j\,u_{2j}^2)]}. \quad \text{(A2)}$$

We now develop a means of incorporating correlations into $P(x, y)$. We wish to consider the probability density of $P(x, y)$ projected onto the variable $(x - y)$, by changing variables to $\xi = x + y$ and $\zeta = x - y$, and integrating over ξ. In the absence of correlations between x and y, this gives a probability density which is the convolution of $P_1(x)$ and $P_2(y)$, centred at $\zeta = (m_1 - m_2)$ rather than at $\zeta = 0$.* For normal distributions with correlations between x and y (with global correlation coefficient r), the probability distribution in ζ is $P_p(\zeta) = \exp\{[-\zeta - (m_1 - m_2)]^2/(2\,u_p^2)\}/[\sqrt{(2\pi)}\,u_p]$, where $u_p = \sqrt{(u_1^2 + u_2^2 - 2\,r u_1\,u_2)}$. Note that the distribution is explicitly centred at $\zeta = (m_1 - m_2)$ rather than at $\zeta = 0$. The results of this integration along lines of constant $(x - y)$ are illustrated in Figure 1 as $P_3(z)$, where the variable z is $(x - y)/\sqrt{2}$ is used so that it has the same scale in Figure 2 as the diagonal in the two-dimensional plot of $P(x, y)$ – note that $P_3(z)\,\mathrm{d}z = P_p(\zeta)\,\mathrm{d}\zeta$.

The confidence level associated with a subsequent similar measurement pair (x, y) having its value of $(x - y)$ in a range of ζ from ζ_1 to ζ_2 is just the integral of the probability density $P_p(\zeta)\,\mathrm{d}\zeta$ from ζ_1 to ζ_2.

The commonest type of confidence interval is not appropriate: one symmetrically situated about the centroid of the known bias, estimated as $m_1 - m_2$. This type is only appropriate when the measured value of $m_1 - m_2$ is to be subtracted out in any subsequent comparison (that is, when y is to be compared with $x - (m_1 - m_2)$). This procedure is not appropriate when the measurements are hypothesized to be centred elsewhere, as is the case for many clients dealing with equivalence statements: their preconception is that the measurements made at the two laboratories should give the same result, and we believe that they would be best served by a simple, symmetric interval about this preconception. Thus we are led to integrating symmetrically about $\zeta = 0$, to find the confidence level C of a result being in the interval $\zeta = -d_C$ to $\zeta = +d_C$. For normal distributions the confidence level is expressible in terms of the cumulative distribution

$$h(x) = 1/\sqrt{(2\,\pi)} \int_{-\infty}^{x} \exp(-X^2/2)\,\mathrm{d}X,$$

where

$$h(x) = 1/2\,\mathrm{erf}(x/\sqrt{2}) = (1/2) \int_0^{x/\sqrt{2}} \exp(-t^2)\,\mathrm{d}t,$$

$$\begin{aligned} C = \int_{-d_C}^{+d_C} P_p(\zeta)\,\mathrm{d}\zeta = \\ h[(m_1 - m_2 + d_C)/u_p] - \\ h[(m_1 - m_2 - d_C)/u_p] = \\ 1/2\,(\mathrm{erf}\{[d_C - (\,|\,m_2 - m_1\,|\,)]/(u_p\sqrt{2})\} + \\ \mathrm{erf}\{[d_C + (\,|\,m_2 - m_1\,|\,)]/(u_p\sqrt{2})\}), \end{aligned} \quad \text{(A3)}$$

which can be solved iteratively for d_C for any particular value of C.

More general forms of $P(x, y)$ can be simply handled by a two-dimensional integration, integrating over the region from $y = x - d$ to $y = x + d$, as shown in Figure 2. Changing variables from x, y to $\xi = x + y$ and $\zeta = x - y$, with Jacobian $-1/2$, the integral in this region over the probability density $P(x, y)$ is

$$\begin{aligned} &\int_{x=-\infty}^{x=+\infty} \int_{y=x-d}^{y=x+d} P(x, y)\,\mathrm{d}y\,\mathrm{d}x = \\ &\int_{\xi=-\infty}^{\xi=+\infty} \int_{\zeta=+d}^{\zeta=-d} \times \\ &P[(\xi + \zeta)/2,\ (\xi - \zeta)/2]\,(-1/2)\,\mathrm{d}\zeta\,\mathrm{d}\xi = \\ &\int_{\zeta=-d}^{\zeta=+d} \int_{\xi=-\infty}^{\xi=+\infty} \times \\ &P[(\xi + \zeta)/2,\ (\xi - \zeta)/2]\,(1/2)\,\mathrm{d}\xi\,\mathrm{d}\zeta. \end{aligned} \quad \text{(A4)}$$

This is the general form which can be used with arbitrary probability density functions, and can include the effects of correlations between x and y that go beyond a simple convolution.

* This distinction is crucial, since although the best estimate of ζ is $m_1 - m_2$, it is *not* the estimate normally contemplated in equivalence statements, which retain the a priori estimate that $\zeta = 0$.

Appendix B

Degrees of freedom

For most applications in metrology the degrees of freedom has been a parameter which we have been able to neglect. However, any method purporting to quantify the level of agreement between national measurement laboratories needs to be capable of rigorously handling the degrees of freedom. Fortunately, the *Guide* is clear enough to need only careful interpretation to evaluate the confidence interval for agreement including the effect of degrees of freedom, and we present our interpretation here.

The *Guide* recommends the use of degrees of freedom, ν, to describe a limited knowledge of the standard uncertainty of a set of measurements. Its use is explicitly recommended for any uncertainty component evaluated by repeated measurement (simply the number of independent measurements minus the number of independent fitted parameters). For an uncertainty component evaluated in other ways, the degrees of freedom may be estimated from the uncertainty Δu in the uncertainty u: $\nu \approx 0.5\,(\Delta u/u)^{-2}$. The effective degrees of freedom, ν_{eff}, of all uncertainty components is evaluated by the Welch-Satterthwaite formula:

$$\nu_{\text{eff}} = \left(\Sigma_i\, u_i^2\right)^2 / \left(\Sigma_i\, u_i^4/\nu_I\right).$$

The probability density used to describe a measurement is a Student's t-distribution with ν_{eff} degrees of freedom.

The *Guide* refers the reader to standard textbooks on statistics concerning the Student's t-distribution for ν degrees of freedom. This distribution applies to any variable which is the ratio of two random variables [8], with the numerator drawn from a normal distribution, and the denominator drawn from an independent random variable which is the square root of a random variable distributed as χ^2 with ν degrees of freedom. For a sample of $\nu + 1$ independent measurements that are normally distributed, the ratio of the deviation of the sample mean from the "true" mean to the sample variance divided by $\sqrt{(\nu+1)}$ may be shown to have a Student's t-distribution, and it is in this pure Type A form that the Student's t-distribution is most often employed in metrology. If the variance is estimated in any other independent way, by pooling results or through Type B techniques, it is also possible to show that the Student's t-distribution should apply. With Type B methods the major concern usually will be over the adequacy of the χ^2 distribution as a description of the variation of the uncertainty in the uncertainty.

For comparing measurement pairs, the uncertainties are combined as $u_p = \sqrt{(u_1^2 + u_2^2)}$, deliberately omitting fully correlated $(r_i = +1)$ uncertainty terms in assembling $u_1 = \sqrt{\left(\Sigma_i\, u_{i,1}^2\right)}$ and $u_2 = \sqrt{\left(\Sigma_i\, u_{i,2}^2\right)}$, and the degrees of freedom $\nu_{\text{eff},1} = (\Sigma_i\, u_{i,1}^4/\nu_{i,1})/(\Sigma_i\, u_{i,1}^2)^2$ and $\nu_{\text{eff},2} = (\Sigma_i\, u_{i,2}^4/\nu_{i,2})/(\Sigma_i\, u_{i,2}^2)^2$. The overall degrees of freedom for the comparison is $\nu_{\text{eff},p} = (u_1^2 + u_2^2)^2/(u_1^4/\nu_{\text{eff},1} + u_2^4/\nu_{\text{eff},2})$, and the distribution of the differences will be approximated by a Student's t-distribution with $\nu_{\text{eff},p}$ degrees of freedom. It is not necessary to truncate the degrees of freedom to an integer. The probability distribution, P_p, is again given by the convolution of P_1 and P_2, which is approximately a Student's t-distribution with mean $(m_2 - m_1)$, standard deviation $u_p = \sqrt{(u_1^2 + u_2^2)}$ and effective degrees of freedom $\nu_{\text{eff},p}$. As an additional note, we believe that an estimate of the degrees of freedom of the comparison, assembled in this way, should be tested against the metrologists' intuitive appreciation of the uncertainty in the uncertainty: $\nu_{\text{eff},p} \approx 0.5\,(\Delta u_p/u_p)^{-2}$. In general we would also advocate taking the smaller of these two estimates of the degrees of freedom of the comparison. The effective degrees of freedom for the measurement pair is given by $\nu_{\text{eff},p}$ which we simply refer to as ν.

With the above based closely on the *Guide*, the only new element which we have to introduce is the asymmetric integration of the Student's t-distribution: its probability density function is $f(t) = \Gamma[(\nu+1)/2]/(1 + t^2/\nu)^{[(\nu+1)/2]}\,\Gamma(\nu/2)\sqrt{(\pi\nu)}$, and its cumulative distribution is $g(T) = \int_{-\infty}^{T} f(t)\,\mathrm{d}t$ for a confidence level C, forming a confidence interval $(-d_C, +d_C)$ centred on the assumption that the means are equal. Thus we solve for d_C in the integral equation

$$C = \int_{-d_C}^{+d_C} P_p(t)\,\mathrm{d}t = g[(m_1 - m_2 + d_C)/u_p] - g[(m_1 - m_2 - d_C)/u_p]. \quad \text{(B1)}$$

The confidence interval, $d_{0.95}$, is obtained by symmetric integration of P_p about 0, until the 0.95 confidence level is achieved. Table 1 and Figure 3 illustrate the exact solutions of (B1) for a confidence level of 95 %, and show the dependence of $d_{0.95}$ with u_p, $|\,m_2 - m_1\,|$, and ν.

A simple, but slightly restricted, numerical approximation of $d_{0.95}$ including the effective degrees of freedom ν is given by

$$d_{0.95} \approx |\,m_2 - m_1\,| + a\,\{1.645 + 0.3295 \times \exp[-4.05\,(|\,m_2 - m_1\,|/u_p)]\}\,u_p, \quad \text{(B2)}$$

where

$$a = 0.283 + 0.717\,b + 0.042\,b^3 \exp[-0.399\,(|\,m_2 - m_1\,|/u_p)^2]$$

and

$$b = 1.960 - 3.162/v + 5.46/(\nu - 0.607).$$

Table 1. Quantified equivalence for 95 % confidence, including degrees of freedom. A measurement pair with means m_1 and m_2; with a combined standard uncertainty u_p; and effective degrees of freedom, ν, gives the confidence interval, $(-d_C, +d_C)$. The values of d_C/u_p are tabulated with respect to the normalized difference $|m_2 - m_1|/u_p$ and the effective degrees of freedom ν = 1, 2, 3, 4, 5, 6, 8, 10, 14, 25 and ∞.

$\lvert m_2 - m_1 \rvert / u_p$	$\nu = \infty$	$\nu = 25$	$\nu = 14$	$\nu = 10$	$\nu = 8$	$\nu = 6$	$\nu = 5$	$\nu = 4$	$\nu = 3$	$\nu = 2$	$\nu = 1$
0.0	1.96	2.06	2.14	2.23	2.31	2.45	2.57	2.78	3.18	4.30	12.71
0.2	2.00	2.10	2.18	2.26	2.34	2.47	2.60	2.80	3.20	4.31	12.71
0.5	2.18	2.27	2.34	2.41	2.48	2.61	2.72	2.92	3.30	4.38	12.73
1.0	2.65	2.71	2.77	2.83	2.89	3.00	3.09	3.26	3.60	4.59	12.78
1.5	3.15	3.21	3.26	3.32	3.37	3.46	3.55	3.69	3.99	4.91	12.88
2.0	3.65	3.71	3.76	3.81	3.86	3.95	4.03	4.16	4.44	5.28	13.01
2.5	4.15	4.21	4.26	4.31	4.36	4.45	4.52	4.65	4.91	5.70	13.18
3.0	4.65	4.71	4.76	4.81	4.86	4.94	5.02	5.14	5.39	6.14	13.38
3.5	5.15	5.21	5.26	5.31	5.36	5.44	5.52	5.64	5.88	6.60	13.60
4.0	5.65	5.71	5.76	5.81	5.86	5.94	6.02	6.14	6.37	7.07	13.85
5.0	6.65	6.71	6.76	6.81	6.86	6.94	7.02	7.13	7.37	8.02	14.43
7.5	9.15	9.21	9.26	9.31	9.36	9.44	9.52	9.63	9.86	10.47	16.17
10.0	11.64	11.71	11.76	11.81	11.86	11.94	12.02	12.13	12.36	12.95	18.18

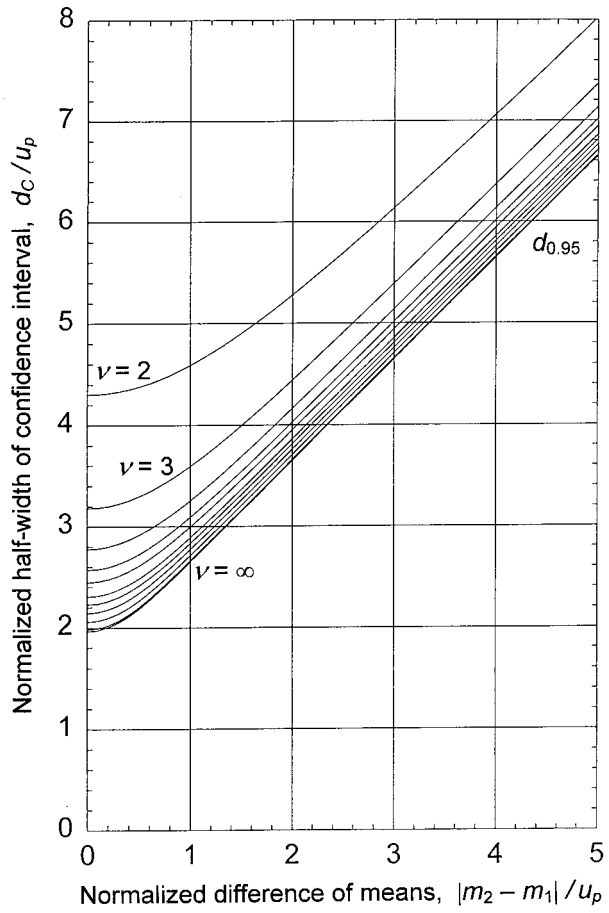

Figure 3. Quantified equivalence for 95 % confidence, including degrees of freedom. A measurement pair with means m_1 and m_2; with a combined standard uncertainty u_p; and effective degrees of freedom, ν, gives the confidence interval, $(-d_C, +d_C)$. The variation of d_C/u_p with normalized difference $|m_2 - m_1|/u_p$ is plotted for effective degrees of freedom ν = 2, 3, 4, 5, 6, 8, 10, 14, 25, 100 and ∞.

The term b is an approximation of the ratio of the breadths of the 95 % confidence interval from the Student's t-distributions with ν and infinite degrees of freedom. This numerical approximation of $d_{0.95}$ is accurate to within 6 % for $\nu \geq 2$ in Table 1, and should be sufficient for all practical purposes. It has the added benefit of interpolating conveniently for non-integer degrees of freedom. Equation (B2) is simple enough be evaluated with a hand calculator or spreadsheet program; it requires no integration or iterative procedures; and it retains the same form as (2).

References

1. *Guide to the Expression of Uncertainty in Measurement*, Geneva, International Organization for Standardization, 1993.
2. Taylor B. N., Kuyatt C. E., *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1297, 1994.
3. *The Expression of Uncertainty and Confidence in Measurement for Calibrations*, NIS 3003, Edition 8, May 1995, NAMAS Executive.
4. *Proficiency testing by interlaboratory comparisons*, ISO Guide 43-1 and Guide 43-2, Geneva, International Organization for Standardization, 1996.
5. *Guidelines for the organization of comparisons*, EUROMET Guidance Document #3, available as DFM-1997-R20, from Danish Institute of Fundamental Metrology, Lyngby, Denmark, 1997.
6. *Mutual Recognition of Calibration Services of National Metrology Institutes*, NORAMET Document #8, available from National Research Council of Canada, 1998.
7. Taylor B. N., NIST, personal communication, December 1997.
8. Freund J. E., *Mathematical Statistics*, Englewood Cliffs, New Jersey, Prentice Hall, 1971.

Received on 11 December 1997 and in revised form on 4 May 1998.