Transfer Standard Uncertainty Can Cause Inconclusive Inter-Laboratory Comparisons

John Wright, Blaza Toman: National Institute of Standards and Technology (NIST)
Bodo Mickan, Gerd Wübbeler, Olha Bodnar and Clemens Elster: Physikalisch-Technische Bundesanstalt (PTB)
Corresponding Author: john.wright@nist.gov

Abstract

Inter-laboratory comparisons use the best available transfer standards to check the participants' uncertainty analyses, identify underestimated uncertainty claims or unknown measurement biases, and improve the global measurement system. For some measurands, instability of the transfer standard can lead to an inconclusive comparison result. If the transfer standard uncertainty is large relative to a participating laboratory's uncertainty, the commonly used standardized degree of equivalence ≤ 1 criterion does not always correctly assess whether a participant is working within their uncertainty claims. We show comparison results that demonstrate this issue and propose several criteria for assessing a comparison result as passing, failing, or inconclusive. We investigate the behavior of the standardized degree of equivalence and alternative comparison measures for a range of values of the transfer standard uncertainty relative to the individual laboratory uncertainty values. The proposed alternative criteria successfully discerned between passing, failing, and inconclusive comparison results for the cases we examined.

1. Introduction

Under the direction of the Comité International des Poids et Mesures (CIPM) and the Mutual Recognition Arrangement [1], committees are working to (1) facilitate the assembly and approval of Calibration and Measurement Capabilities (CMCs) for member National Metrology Institutes (NMIs), and (2) conduct laboratory comparisons that can be used to assess the validity and improve the CMCs. More than 1000 comparisons are listed in the Key Comparison Database [2] and the methodology for conducting and processing a comparison has advanced [3, 4, 5, 6]. But using the results of comparisons to accept or reject a stated capability is not a simple decision and there is still work to be done to make CMC approval a more objective and reliable process. In this paper we will use comparison data to illustrate problems introduced by large transfer standard uncertainty and propose criteria to decide whether a participant's results are equivalent ("passing"), not equivalent ("failing"), or inconclusive.

2. WGFF Guidelines for CMC Uncertainty

In 2013, the Working Group for Fluid Flow (WGFF) produced the WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty [7], an effort to have NMIs use a common approach and terminology in their CMC statements. The Guidelines state that the CMC standard uncertainty (u_{CMC}) is composed of: (1) a Type B base uncertainty (u_{base}) of the laboratory's reference standard obtained by using the law of propagation of uncertainty as described in the Guide to the Expression of Uncertainty in Measurement (GUM) [8] and (2) a Type A uncertainty based on n calibration results measured using the Best Existing Device (BED), i.e.,

$$u_{\mathrm{CMC}} = \sqrt{u_{\mathrm{base}}^2 + \left(\frac{s}{\sqrt{n}}\right)_{\mathrm{BED}}^2}\,. \qquad (1)$$

The standard uncertainty u_{CMC} is multiplied by a coverage factor $k = 2$ to obtain the expanded uncertainty listed in the BIPM's Key Comparison Database. The quantity $s/\sqrt{n}$ is the standard deviation of the mean where s is the sample standard deviation and n is the number of repeated measurements. Inclusion of this Type A component in u_{CMC} is called for by the CIPM [9] and International Laboratory Accreditation Cooperation [10]. It covers sources of uncertainty that are not yet known by the lab [11] and hence not yet accounted for by u_{base}. Note that the standard deviation of the mean is the appropriate measure of the Type A uncertainty when reporting the average of n measurements, not the sample standard deviation s.[1] The values of the reference standard (x_1) and of the inherent bias of the BED prior to calibration (0) that are associated with the uncertainties in Eqn. 1 are represented by normal, Gaussian probability density functions (PDF's), $\mathrm{N}(\mu, \sigma)$ in Figure 1 where μ is the mean and σ is the standard uncertainty.

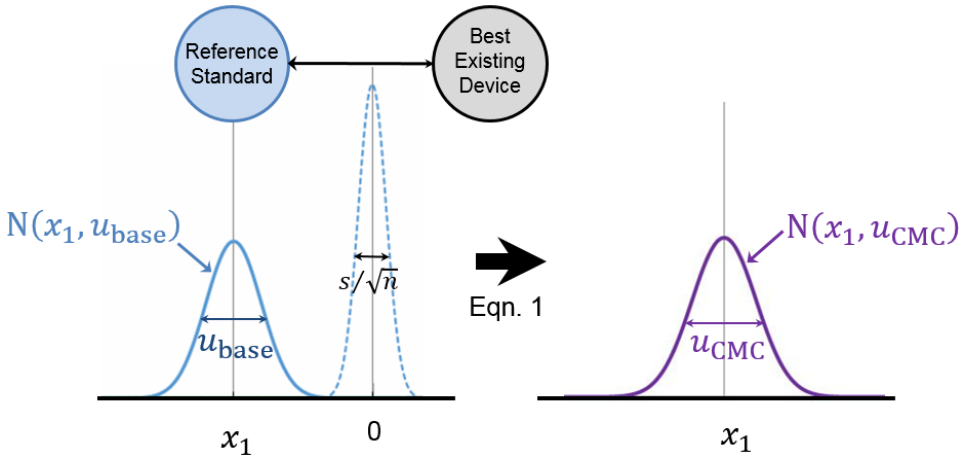


Figure 1. Probability density functions for the value of the reference standard, with standard uncertainty u_{base} (Type B), for the inherent bias of the BED prior to calibration, with standard uncertainty $s/\sqrt{n}$ (Type A), and the calibrated output of the BED, with standard uncertainty u_{CMC}. $u_{\mathrm{CMC}}$is the CMC standard uncertainty that will be included in the Key Comparison Database.

The Type B uncertainty of the laboratory's reference standard u_{base} is independent of the particular device being calibrated and is a critical input to a laboratory comparison [7]. During analysis of comparison results, it is combined with transfer standard uncertainties and the Type A uncertainties measured during the comparison to arrive at the uncertainty of each participant's reported value (Eqn. 6).

3. Review of Inter-Laboratory Comparisons and Transfer Standard Uncertainty

[1] Note that someone interested in the Type A uncertainty of a *single* measurement rather than the average of n measurements should use s, not $s/\sqrt{n}$.

To verify the calibration and measurement capabilities of laboratories, a working group selects a Pilot lab and conducts an inter-laboratory comparison. The Pilot lab ships one or more transfer standards between a set of participating labs and the results from each lab are used to calculate a comparison reference value (CRV). The difference between each participant and the CRV (the degree of equivalence, $d_i = x_i - x_{\mathrm{CRV}}$) is used to assess whether participants are meeting their uncertainty claims and provides an important basis for approval, disapproval, or modification of CMCs. The comparison also allows labs to validate their largely paper-based uncertainty analysis with experimental data. There is a commonly applied system for assessing comparison results, called herein "Criterion A" ($|En_i| \leq 1$).

It is important to quantify the uncertainty of the transfer standard used in a comparison. The standard uncertainty of the transfer standard, u_{TS} should account for the calibration drift of the transfer standard (and its associated instrumentation) during the comparison, temperature sensitivities, pressure sensitivities, property sensitivities, and perhaps other components specific to the transfer standard:

$$u_{\mathrm{TS}} = \sqrt{u_{\mathrm{drift}}^2 + u_{\mathrm{T}}^2 + u_{\mathrm{P}}^2 + u_{\mathrm{prop}}^2 + \cdots}. \tag{2}$$

The uncertainties in Eqn. 2 should be quantified during preliminary tests organized by the Pilot lab. In flow comparisons, the uncertainty due to calibration drift is usually the largest component. It can be quantified by performing repeated calibrations in the Pilot lab using the same reference standard before, during, and immediately after the comparison as shown in Figure 2. Linear transfer standard drift over time can be corrected, but for many transfer standards, the drift is effectively random. Unfortunately, the uncertainty of the transfer standard is often not known until the conclusion of the comparison when repeated calibrations to quantify long-term drift are complete. The uncertainty due to drift of the transfer standard over the course of a comparison can be estimated by

$$u_{\mathrm{drift}} = s, \tag{3}$$

where s denotes the standard deviation of the n measurements at each flow set point. For small n, e.g. $n < 5$, a rectangular distribution can be applied to the range of the calibration changes observed by the Pilot lab, $(\varepsilon_{\max} - \varepsilon_{\min})$, i.e. $u_{\mathrm{drift}} = \frac{(\varepsilon_{\max} - \varepsilon_{\min})}{2\sqrt{3}}$.

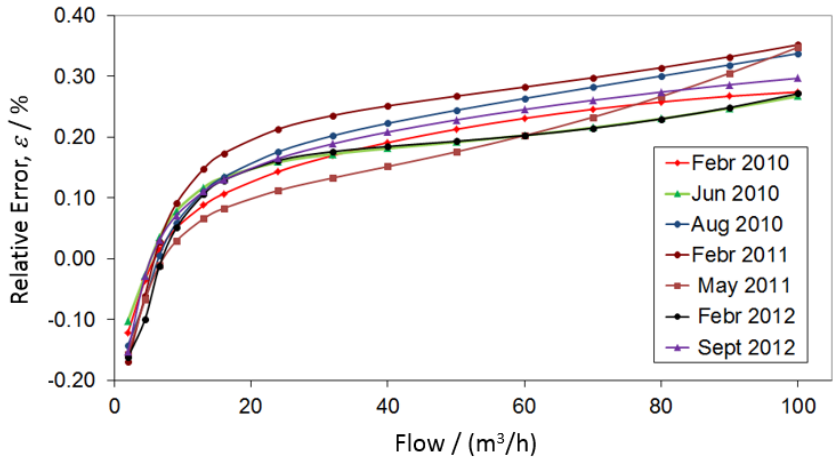


Figure 2. An example of Pilot lab testing of a transfer standard to quantify u_{drift} from reference [12]. In this case, Eqn. 3 was applied, giving $u_{\text{drift}} = 0.04$ %. Note that a single value can be used here since the values of u_{drift} are essentially constant over all relevant flow rates.

The uncertainty-weighting methods published by Cox [3] are often used to calculate the CRV, x_{CRV}, its uncertainty, $u_{x_{\text{CRV}}}$, the degrees of equivalence, $d_i = x_i - x_{\text{CRV}}$, and the uncertainty of the degree of equivalence, u_{d_i}. When carried out with independent laboratories results (i.e., data are consistent and no covariance between participants) Cox's approach uses the following equations to estimate the uncertainties, $u_{x_{\text{CRV}}}$, and u_{d_i}

$$\frac{1}{{u_{x_{\text{CRV}}}}^2} = \frac{1}{u_{x_1}^2} + \frac{1}{u_{x_2}^2} + \cdots + \frac{1}{u_{x_n}^2} \text{ and} \tag{4}$$

$$u_{d_i} = \sqrt{u_{x_i}^2 - u_{x_{\text{CRV}}}^2}. \tag{5}$$

However, the uncertainty of the reported value (called u_{x_i} by Cox) is ***not*** simply the uncertainty of the participant's flow reference ($u_{\text{base}\,i}$): it must also include uncertainties introduced by the transfer standard and the repeatability of the reported value at each set point. These extra uncertainty components are often significant relative to the participating labs' base uncertainties. The uncertainty of the reported value is:

$$u_{x_i} = \sqrt{u_{\text{base}\,i}^2 + u_{\text{TS}}^2 + \left(\frac{s_i}{\sqrt{n}}\right)^2} = \sqrt{u_{\text{CMC}\,i}^2 - \left(\frac{s_i}{\sqrt{n}}\right)_{\text{BED}}^2 + u_{\text{TS}}^2 + \left(\frac{s_i}{\sqrt{n}}\right)^2} \tag{6}$$

The $s_i/\sqrt{n}$ term is the standard deviation of the mean of the n measurements made at each flow set point during the comparison and quantifies the reproducibility of the measurements made in the participant's lab. While the CMC uncertainty uses the Type A uncertainty measured with the best existing device, the uncertainty of the reported value in a comparison should use the Type A uncertainty for the transfer standard.

Therefore, it is necessary for participants to report $u_{\text{base }i}$ to the Pilot lab or, alternatively, $\left(s_i/\sqrt{n}\right)_{\text{BED}}$ so that $u_{\text{base }i}$ can be calculated from u_{CMC}, as shown in Eqn. 6.

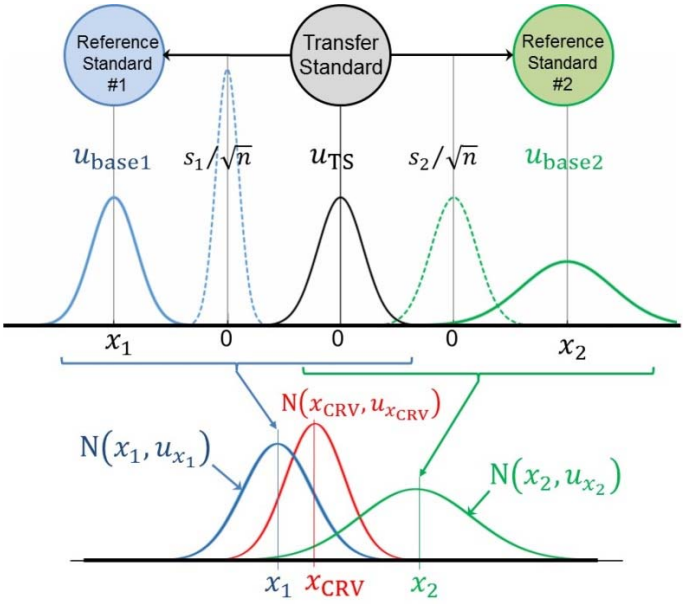


Figure 3. Probability density functions for a bi-lateral comparison. In this example, $u_{\text{TS}}/u_{\text{base 1}} = 1$, $u_{\text{TS}}/u_{\text{base 2}} = 0.5$.

Figure 3 illustrates the processing of bi-lateral comparison data via Eqns. 4 and 6. The Type A component from the best existing device that was used to calculate u_{CMC} is not used during the comparison data processing. Instead, the Type A uncertainty obtained by the participant during the collection of the comparison data is used.

Alternatives to Cox's methods for calculating the CRV and its uncertainty are available [13, 14, 15] and they may be more appropriate for some comparisons. In some comparisons, multiple methods have been applied and presented in the comparison reports [16] and the relatively small differences between the CRVs and their uncertainties give increased confidence in the comparison results. In other cases, it is important to be aware of the differences between CRV calculation methods and to use the most appropriate method.

4. Presently Used Comparison Pass / Fail Criterion

In 2013, the CIPM requested that Pilot labs give clearer guidance to CMC reviewers as to whether or not a comparison supports a participant's CMC uncertainty. Many comparisons have used the standardized degree of equivalence,

$$En_i = \frac{d_i}{2u_{d_i}}, \tag{7}$$

and what we will call:

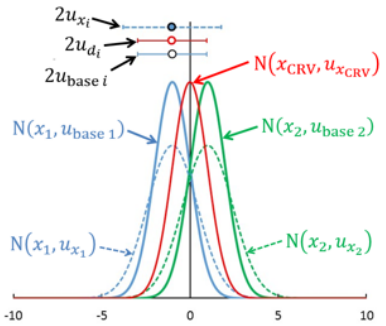
Criterion A: Participant i passes if $\lvert En_i \rvert \leq 1$ and fails if $\lvert En_i \rvert > 1$.

Some Pilots have added a "warning" (not failing) level if $|En_i|$ is between 1 and 1.2. Unfortunately, a transfer standard uncertainty u_{TS} that is large relative to a participating lab's uncertainty $u_{\mathrm{base}\,i}$ leads to inconclusive comparison results, even when $|En_i| \leq 1$. Large u_{TS} leads to large u_{x_i} and large u_{d_i} and can hence result in $|En_i| \leq 1$ even when $u_{\mathrm{base}\,i}$ is severely underestimated. Some graphical examples from a fictitious "bi-lateral comparison example" help to explain the deficiencies of Criterion A.

In the first bi-lateral comparison example (Figure 4a), the reported values from the two participants are $x_1 =$ -1 and $x_2 = 1$, both labs have the same uncertainty for their reference standards, i.e., $u_{\mathrm{base}\,1} = u_{\mathrm{base}\,2} = 1$, and the transfer standard also has uncertainty of 1. Neglecting the repeatability component, the uncertainty of each participant's reported value is the combined standard uncertainty accounting for u_{TS} and $u_{\mathrm{base}\,i}$, in this case the same value for both labs, $u_{x_1} = u_{x_2} = 1.414$. The comparison reference value $x_{\mathrm{CRV}} = 0.0$ and the uncertainty of the CRV, $u_{x_{\mathrm{CRV}}} = 1$. Finally, the standardized degree of equivalence for the two labs $|En_1| = |En_2| = 0.5$, i.e. equivalent by the $|En_i| \leq 1$ criterion.

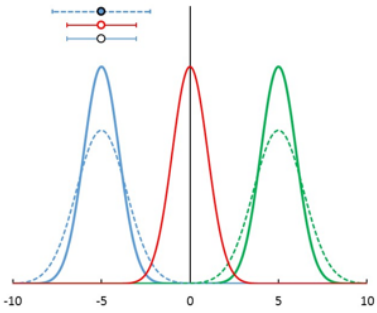
Figure 4a tabulates the quantities for this example and uses a format to present the comparison results that we will use throughout this paper. Figure 4a plots Gaussian probability density functions (PDFs) for the participants' reported values and the CRV. Two versions of the participants' PDFs are shown, one for the lab's base uncertainty $u_{\mathrm{base}\,i}$ (solid lines), and a second that uses the uncertainty of the reported value u_{x_i} (dashed lines). In Figure 4a, the high degree of overlap of all the PDFs is a strong indication of equivalence between the two labs and the CRV.

a) Equivalent (✓)



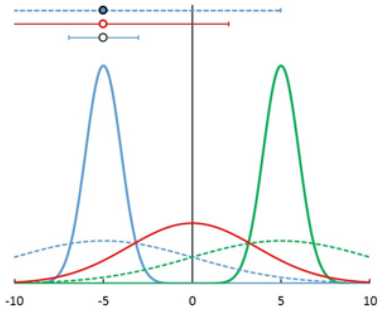
	Lab 1	Lab 2	x_{CRV}
x_i	-1.000	1.000	0.000
$u_{\mathrm{base}\,i}$	1.000	1.000	
u_{TS}	1.000	1.000	
$s_i/\sqrt{n}$	0.000	0.000	
$u_{x\,i}$	1.414	1.414	1.000
$u_{d\,i}$	1.000	1.000	
En_i	-0.500	0.500	

b) Non-Equivalent (×)



Lab 1	Lab 2	x_{CRV}
-5.000	5.000	0.000
1.000	1.000	
1.000	1.000	
0.000	0.000	
1.414	1.414	1.000
1.000	1.000	
-2.500	2.500	

c) Inconclusive (?)



Lab 1	Lab 2	x_{CRV}
-5.000	5.000	0.000
1.000	1.000	
5.000	5.000	
0.000	0.000	
5.099	5.099	3.606
3.606	3.606	
-0.693	0.693	

Figure 4. a) Clear Equivalence: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 1$, $|En_1| = |En_2| = 0.5$, and the participating labs are equivalent. The dashed curves represent the probability density functions for u_{x_i}, i.e. including the transfer standard uncertainty. **b)** Clear non-equivalence: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 1$, $|En_1| = |En_2| = 2$, and the participating labs are not equivalent. **c)** Inconclusive: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 5$, $|En_1| = |En_2| = 0.69$ (≤ 1), but the participating labs reported results do not appear to be equivalent.

We can use the mean and the expanded uncertainty of a participating lab ($2u_{\mathrm{base}\,i}$) to calculate a 95 % uncertainty interval (a_i, b_i) for their measurement where a_i is the 2.5$^{\mathrm{th}}$ percentile of the distribution and b_i is the 97.5$^{\mathrm{th}}$ percentile. Figure 4a has three circular symbols representing x_1 and they have error bars representing the 95 % uncertainty interval based on u_{x_i} (dashed), u_{d_i} (red), and $u_{\mathrm{base}\,i}$ (blue).

In a second example (Figure 4b), the values reported by the two labs are quite different from each other (-5 and 5), $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 1$, and $|En_1| = |En_2| = 2.5$, i.e. by the $|En_i| \leq 1$ criterion, the two labs' results are not equivalent. The lack of overlap of the Lab 1 and Lab 2 PDFs and the various error bars not crossing the CRV at $x = 0$ also indicate that the results are not equivalent.

In our third example (Figure 4c), the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 5$, greatly weakening the ability to discern differences between the two participants (and the explanatory power of the comparison). Despite the large difference in their reported values, $|En_1| = |En_2| =$ 0.69. The observed large difference in the laboratory results should not necessarily be viewed as an indication of non-equivalence here. The difference could indeed be caused by unrecognized laboratory effects, but it may likewise result from instability of the transfer standard. Therefore, a situation like that shown in Figure 4c should be viewed as being inconclusive.

Our review of past key and regional comparison reports shows that when values for u_{TS} are given, $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ is often larger than 1, and in some cases greater than 5. The example in Figure 4c illustrates the inadequacy of the generally used $|En_i| \leq 1$ criterion in cases where the transfer standard uncertainty is large relative to the participants' base uncertainties. Large transfer standard uncertainty broadens the PDF of the reported value (dashed PDF), leads to a large value for the uncertainty of the degree of equivalence (u_{d_i}), and makes it possible to obtain $|En_i| \leq 1$ even when d_i is large relative to the lab's uncertainty claim $u_{\mathrm{base}\,i}$.

The pass / fail decision can be treated as a statistical hypothesis test to check if the unilateral or bi-lateral degree of equivalence is significantly different from zero. This is the approach many comparison report readers visually employ when they look at plots of comparison results: do the 95 % coverage intervals of the degrees of equivalence include zero? In fact, the Mutual Recognition Arrangement documents [17] state that comparison results will be presented as "the deviation from the key comparison reference value and the expanded uncertainty of this deviation computed at a 95 % level of confidence".

Figure 5 presents the degrees of equivalence for a liquid flow comparison [18] in which $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ ranged between 2.2 and 5.7. Two versions of the results are shown for each lab, 1) the red open symbols have $2u_{d_i}$

coverage intervals, and 2) the blue open symbols have coverage intervals equal to $2u_{\text{base } i}$. The $2u_{\text{base } i}$ and $2u_{d_i}$ coverage intervals are shown to illustrate the influence of u_{TS} on the results from Criterion A. For Laboratories 2, 4, 5, 6, 8 and 11, including the large transfer standard uncertainty (and $s_i/\sqrt{n}$) in the analysis makes the difference between their results being considered equivalent or not, i.e. the red $2u_{d_i}$ coverage intervals used for the $|En_i| \leq 1$ criterion cross the CRV value (0) while the blue error bars do not.

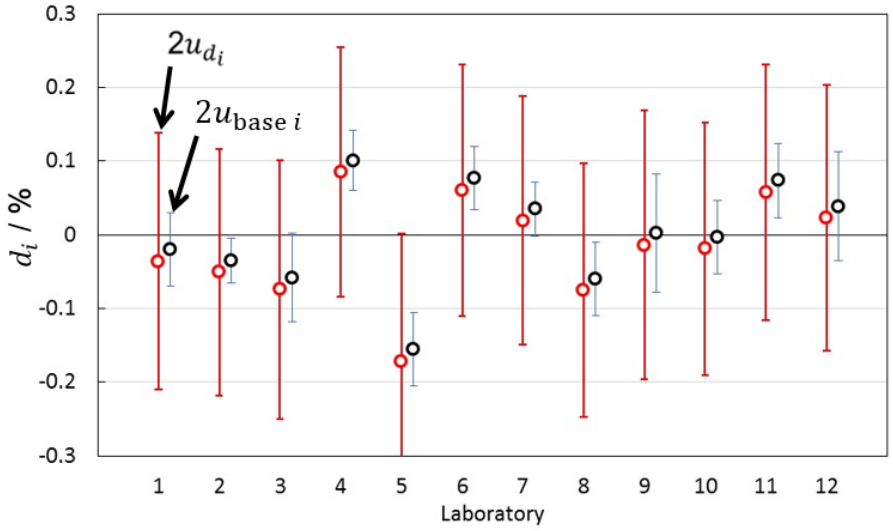


Figure 5. Liquid flow comparison results [18] with two versions of error bars: 1) $2u_{d_i}$ (red), and 2) $2u_{\text{base } i}$ (blue).

5. Behavior of PDFs and En_i over the $d_i/u_{\text{lab } i}$ and $u_{\text{TS}}/u_{\text{base } i}$ Parameter Space

We return to the bi-lateral comparison example described in section 4, *i.e.* $x_1 = -x_2$, $u_{\text{base } 1} = u_{\text{base } 2}$, and negligible $s_i/\sqrt{n}$. Figure 6 plots PDFs for $d_i/u_{\text{base } i}$ and $u_{\text{TS}}/u_{\text{base } i}$ ranging from 1 to 8 and allows us to examine the behavior of $|En_i|$ over the parameter space. For Lab 1, three circular symbols represent x_1 and they have horizontal coverage intervals representing $2u_{x_i}$ (dashed), $2u_{d_i}$ (red), and $2u_{\text{base } i}$ (blue).

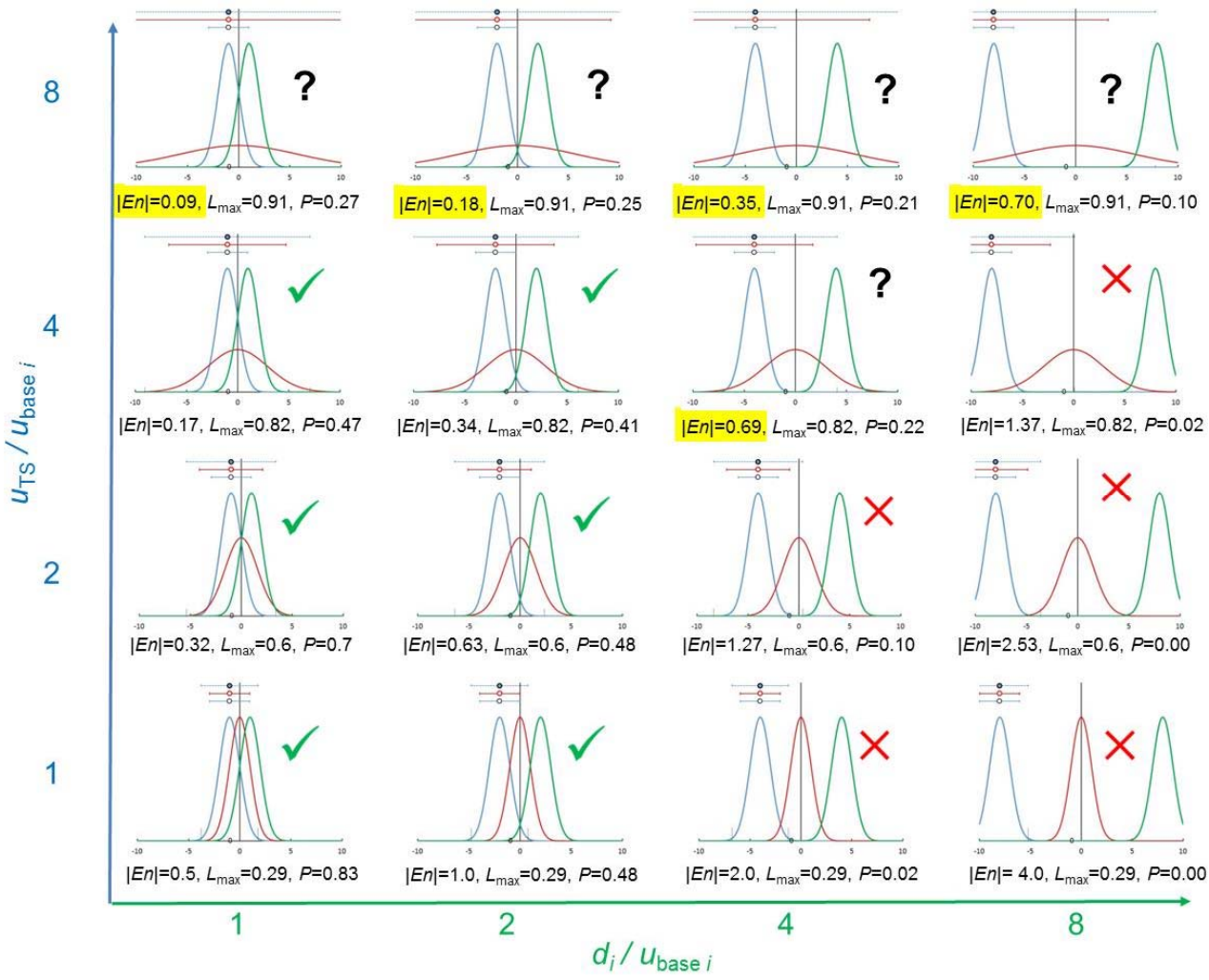


Figure 6. Probability density functions for the bi-lateral comparison example plotted for $d_i/u_{\text{base } i}$ and $u_{\text{TS}}/u_{\text{base } i}$ ranging from 1 to 8 with a visual assessment for ✓ (equivalent or passing), × (not equivalent or failing), or **?** (inconclusive).

We have performed a subjective "visual assessment" of the cases in Figure 6 and assigned the labels ✓ (equivalent or passing), × (not equivalent or failing), or **?** (inconclusive). We have assigned the labels using the following criteria. If $|En_i| > 1$, the reported value is considered not equivalent (×). Neither the $2u_{\text{base } i}$ or $2u_{d_i}$ error bars cover the CRV for the cases marked ×. If the reported value agrees with the CRV within the participant's $2u_{\text{base } i}$ claim (i.e. the solid blue error bars cross 0), we generally considered the reported value equivalent to the CRV (✓). However, when $u_{\text{TS}}/u_{\text{base } i}$ is large, a participant's agreement with the CRV may be due to the transfer standard drifting in a fortuitous way. Hence, cases labelled **?** are considered inconclusive because the uncertainty contributed by the transfer standard is too large.

The southwest quadrant of Figure 6 holds cases where we have strong confidence that the participant's result is equivalent: the participant's reported value is close to the CRV and the uncertainty of the transfer standard

is low. The southeast quadrant holds cases where the differences between the participants are large enough relative to the transfer standard uncertainty that one can clearly decide that the participant is not equivalent.

Five cases in Figure 6 are shaded yellow because they have $|En_i| \leq 1$ (passing Criterion A), but we would visually assess them as inconclusive. Inspection of the error bars and PDFs shows that $|En_i| \leq 1$ may be due to large transfer standard uncertainty (large u_{x_i} and u_{d_i}), not necessarily due to a consistency between the participant's measurement result and the true value of the transfer standard. Note that in a prior publication [19], we proposed passing results in the northwest corner regardless of large transfer standard uncertainty on the grounds that a participant with excellent agreement with the CRV should not be penalized for poor transfer standard performance. In this publication, we recommend an inconclusive result for large transfer standard uncertainty, even when there is excellent agreement between the participant's result and the CRV because that agreement may occur by coincidence.

The WGFF has discussed a possible pass / fail / inconclusive criterion that requires $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ below a threshold value R_{th} to produce a passing result:

> **Criterion B:** Participant i passes if $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq R_{\mathrm{th}}$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant i otherwise.

The $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq R_{\mathrm{th}}$ criterion avoids participants passing solely because the transfer standard uncertainty was large. The value of R_{th} in Criterion B is subjective. Figure 7 shows that as $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ increases (and explanatory power decreases), the $|En_i| \leq 1$ criterion alone passes participants with large differences from the CRV. For example, if $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 4$, the $|En_i| \leq 1$ criterion alone allows a participant with $d_{\mathrm{i}}/u_{\mathrm{base}\,i} = 5.83$ to pass. For $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 2$, a participant with $d_{\mathrm{i}}/u_{\mathrm{base}\,i}$ as large as 3.16 passes. Note that in these cases the laboratory still might have correctly specified its uncertainty and that the large differences to the CRV observed might be caused by fluctuations of the transfer standard. By the definition of En_i and the $|En_i| \leq 1$ criterion, the red u_{d_1} 95 % coverage intervals on x_1 cross the CRV (at $x = 0$) for the sample PDFs shown in Figure 7.

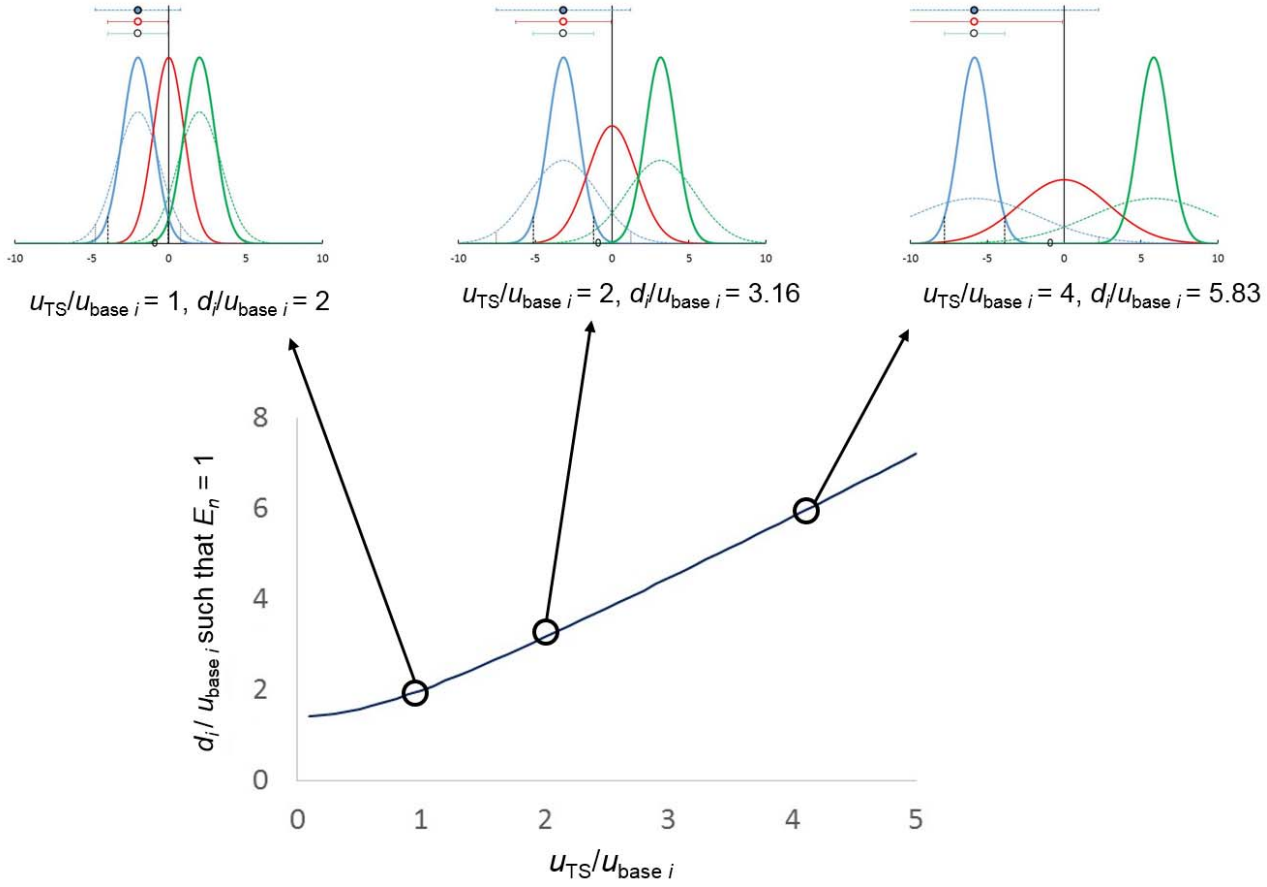


Figure 7. Degree of equivalence that results in $|En_i|$ = 1 versus $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ along with probability density function plots for specific cases of $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ = 1, 2, and 4 for the bi-lateral comparison example.

6. Explanatory Power

Wübbeler *et al.* [20] quantitatively assessed the degradation in the conclusiveness of a key comparison due to transfer standard uncertainty. As proposed in [19], the relevance of a degree of equivalence can be quantitatively assessed by the explanatory power of a hypothesis test that checks whether the degree of equivalence is signifcantly different from zero. The power denotes the probability, prior to carrying out the key comparison, that an underrated uncertainty will be detected through checking the degree of equivalence. Analytical expressions for the power as well as for the loss of power due to instability of the transfer standard are given in [20] and can be used to analyze the scenarios shown in Figure 6. The power results in this work were determined by assuming that the Type A contribution was negligible. Figure 8 shows the loss of power for a bi-lateral comparison for a range of $|\delta_1 - \delta_2|$ values and for the $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values used in Figure 6. δ_i denotes the unknown true laboratory effect which summarizes potentially overlooked effects that have not been accounted for in the uncertainty evaluation of the i-th laboratory. The degree of equivalence d_i can be seen as an estimate of δ_i. As shown in Figure 8, the loss of explanatory power increases monotonically for increasing $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values.

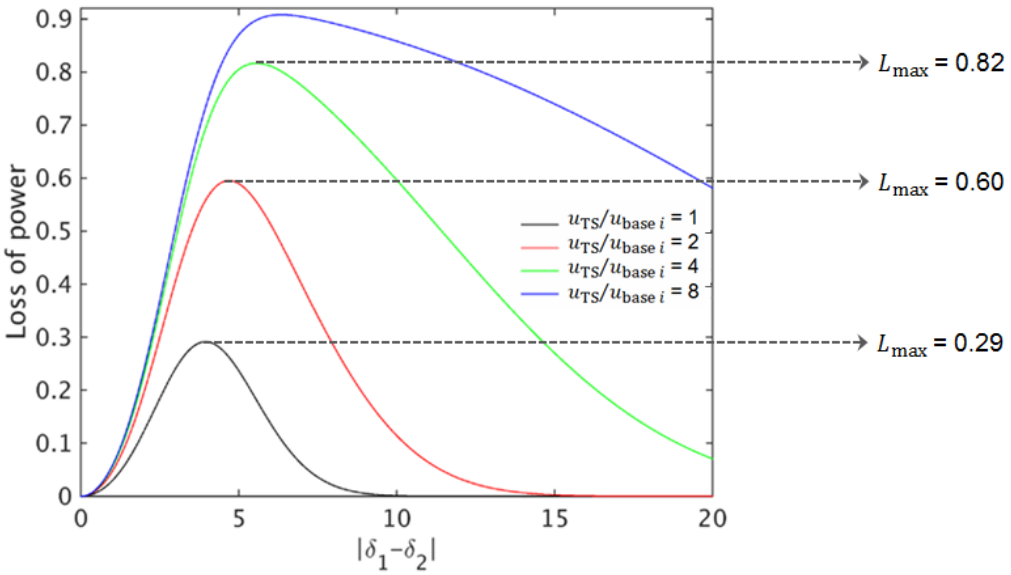


Figure 8. Loss in explanatory power in a bi-lateral comparison as a function of $|\delta_1 - \delta_2|$ for various $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values where the uncertainties quoted by the two laboratories are assumed to be equal. The maximum loss of power ($L_{\max}$) can be used to assess the significance of transfer standard uncertainty on comparison results.

It is useful to quantify the loss of explanatory power introduced by the transfer standard uncertainty. Specifically, one can quantify how the power loss varies over the parameter space of $|\delta_1 - \delta_2|$ and $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ as done in Figure 8. The power loss can also be utilized to design a reliability criterion for comparison results in the presence of an unstable transfer standard. The criterion is based on the maximal loss of power $L_{\max}$, e.g., for $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = 2$ in Figure 8, the maximal loss of power is about 0.6. Setting the maximal tolerable loss of power $L_{\max}$ to a threshold value of L_{th} results in the following criterion:

Criterion C: Participant i passes if the maximal loss of power $L_{\max,i} \leq L_{\mathrm{th}}$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant *i* otherwise.

7. Probability Based Criterion

To derive the probability based criterion of this section we take the view of the introduction to the "Guide to the Expression of Uncertainty in Measurement" [23, section 3.8] that "it is not possible to state how well the essentially unique true value of the measurand is known, but only how well it is believed to be known." This view, applied to the problem at hand interprets each laboratory's reported value x_i and the accompanying uncertainty u_{x_i} as the mean and standard deviation of their belief distribution about the measurand. Also see for example reference [24]. In the absence of other information about the shape of this belief distribution, it can be approximated by a Gaussian probability curve with the same mean and standard deviation. This probability distribution can be used to obtain the expanded uncertainty (1.96 u_{x_i}), or equivalently, a 95% uncertainty interval (a_i, b_i) for the measurand where a_i is the 2.5$^{\mathrm{th}}$ percentile of this Gaussian distribution and b_i is the 97.5$^{\mathrm{th}}$ percentile. The numerical values of the percentiles can be easily obtained using various

software (for instance by using the function NORM.INV in Excel[2]). The usual interpretation of this interval is that the laboratory believes, based on their data and possibly on additional related information, that the measurand lies in the interval (a_i, b_i) with 0.95 probability. In metrology, the additional related information is often referred to as type B evaluation of measurement uncertainty based on manufacturer specifications or reference material certificates. This interpretation of probability, sometimes called subjective, is not based on estimates of relative frequency of the event that the measurand lies in this interval as is the case with frequentist interpretation of probability.

The goal of this article is to propose criteria for the assessment of the labs' uncertainty claims $u_{\text{base } i}$. For this reason, we will use $u_{\text{base } i}$ as the standard deviation of the labs' belief distributions in the calculation of the uncertainty interval, keeping in mind that the probability content (that is, the integrated area of the Gaussian curve between a_i and b_i) is only exactly 0.95 when the transfer standard uncertainty is negligible.

In an interlaboratory comparison, the CRV incorporates results from the n participants, using all of their inputs, and taking into account the uncertainty of the transfer standard. The resulting value x_{CRV} , and the uncertainty $u_{x_{CRV}}$, can be taken to be the mean and standard deviation of a belief distribution for the measurand based on all available information provided by the comparison exercise [21]. Thus it is possible to assess the claims of the individual laboratories by calculating the probability content of their intervals (a_i, b_i) under the Gaussian probability distribution based on the CRV, *i.e.* $\mathrm{N}\left(x_{\mathrm{CRV}}, u_{x_{\mathrm{CRV}}}\right)$. Figure 9 shows this area for Lab 1 as the shaded region, giving the probability labelled P_1. The probability P_i can be calculated in Excel using the following formula:

= NORMDIST(NORM.INV(0.975, x_i, $u_{\text{base } i}$), x_{CRV}, $u_{x_{\mathrm{CRV}}}$, TRUE) –
NORMDIST(NORM.INV(0.025, x_i, $u_{\text{base } i}$), x_{CRV}, $u_{x_{\mathrm{CRV}}}$, TRUE).

If the uncertainty of the transfer standard is small, the content probability of (a_i, b_i) under the CRV based probability distribution should be 0.95 or larger for a well specified initial claim and small content probability discredits the CMC claim. As the transfer standard uncertainty becomes larger, it becomes harder to evaluate the claim, but the content probability of the interval (a_i, b_i) still provides insight which is complementary to the other criteria as is illustrated by Figure 9: as the uncertainty of CRV increases and broadens the PDF of the CRV, P_i decreases, reflecting the reduced confidence in the comparison result. The probability P_i is listed along with $|En_i|$ in Figure 6.

[2] In order to describe materials and procedures adequately, it is occasionally necessary to identify commercial products by manufacturers' name or label. In no instance does such identification imply endorsement by the National Institute of Standards and Technology, nor does it imply that the particular product or equipment is necessarily the best available for the purpose.

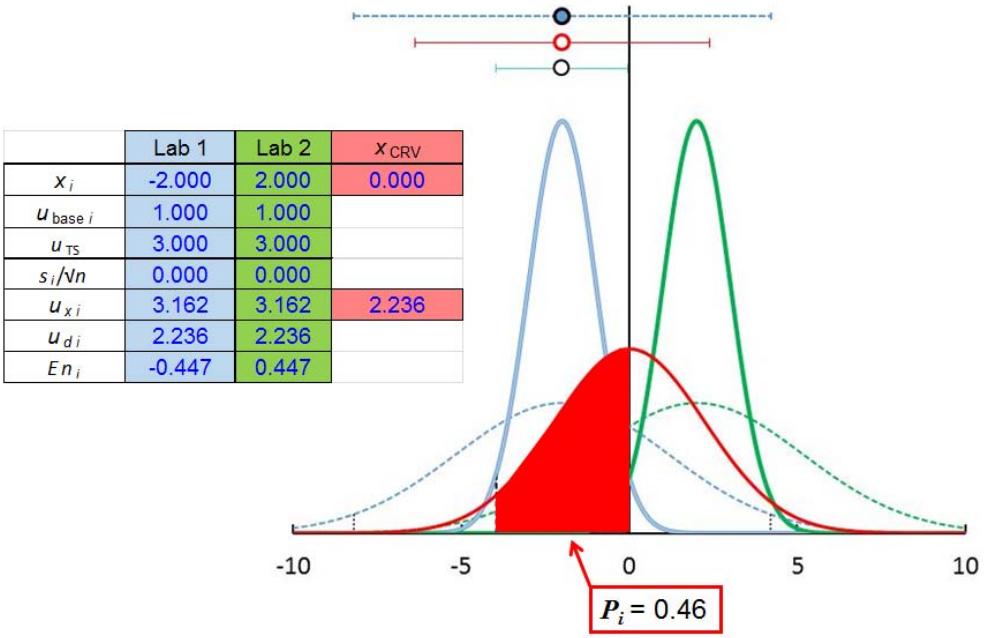


Figure 9. Probability P_i (shaded red) can be utilized for pass / fail criteria.

Under the desirable circumstances when most of the laboratories have reasonably specified uncertainties, and $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ is less than 1, $u_{x_{\mathrm{CRV}}}$ will be smaller than most if not all of the individual $u_{\mathrm{base}\,i}$. In such circumstances, a lab's content probability that is much lower than 0.95 would indicate that the initial claim of the laboratory was not realistic either in its location or in its uncertainty.

When the comparison results are deemed reasonable in the sense that the CRV and its uncertainty are deemed reasonable, it could be useful to have a simple threshold value P_{th} for the probability P_i. Clearly 0.95 or higher is best, but lower values could be judged as acceptable. For some measurands, it may be that a low uncertainty transfer standard does not exist and it may be necessary to account for this by using a value of $P_{\mathrm{th}} < 0.5$. Figure 10 plots P_i versus $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ for cases where $|En_i| = 1$ in the bi-lateral comparison example. To remain consistent with the criterion that $|En_i| \leq 1$ while $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq 1$, P_{th} = 0.48 and to remain consistent with $|En_i| \leq 1$ while $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq 2$, the P_{th} = 0.22. The value of P_{th} used in comparisons may evolve over time as transfer standards improve.

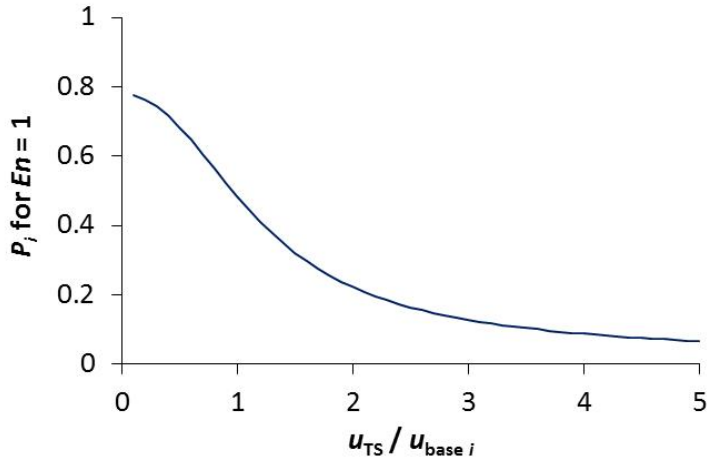


Figure 10. P_i versus $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ for cases where $|En_i| = 1$ in the bi-lateral comparison example.

The standardized degree of equivalence $|En_i|$ and the probability P_i can be used to design a pass / fail / inconclusive criterion that mimics the visual assessments we made in Figure 6. Results in the southeast and northeast quadrants with $|En_i| > 1$ are considered failing. Results with probability P_i below a threshold value P_{th} are inconclusive.

Criterion D: Participant i passes if content probability $P_i \geq P_{\mathrm{th}}$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant i otherwise.

8. Criteria applied to Bi-Lateral Comparison Example and Proposed Threshold Values

Figure 11 shows contour plots for $|En_i|$, $L_{\mathrm{max},i}$, and P_i for $d_i/u_{\mathrm{base}\,i}$ and $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ ranging from 0 to 8. The intention is to show the general behavior of these quantities and use that knowledge to design more sophisticated criteria. The contour plots use green for possible passing values, yellow for possible warning levels, and red for possible values indicating that labs are not equivalent. As described in the discussion of Figure 6, the $|En_i| \leq 1$ criterion is green (passing) for large $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$, giving positive results for regions that should be considered as being inconclusive. In contrast, the content probability P_i and the maximal loss of power $L_{\mathrm{max},i}$ do not suffer from this questionable approval of equivalence for $d_{\mathrm{i}}/u_{\mathrm{base}\,i}$ and $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ > 2 (northeast quadrant). The content probability P_i has the additional feature that it has reduced values in the northwest quadrant. Low content probability P_i occurs when 1) there is poor coincidence between the lab and CRV PDFs (southeast corner), 2) the CRV PDF is broad due to large $u_{x_{\mathrm{CRV}}}$, (large u_{TS}, northwest corner), or 3) when both of these conditions apply (northeast quadrant).

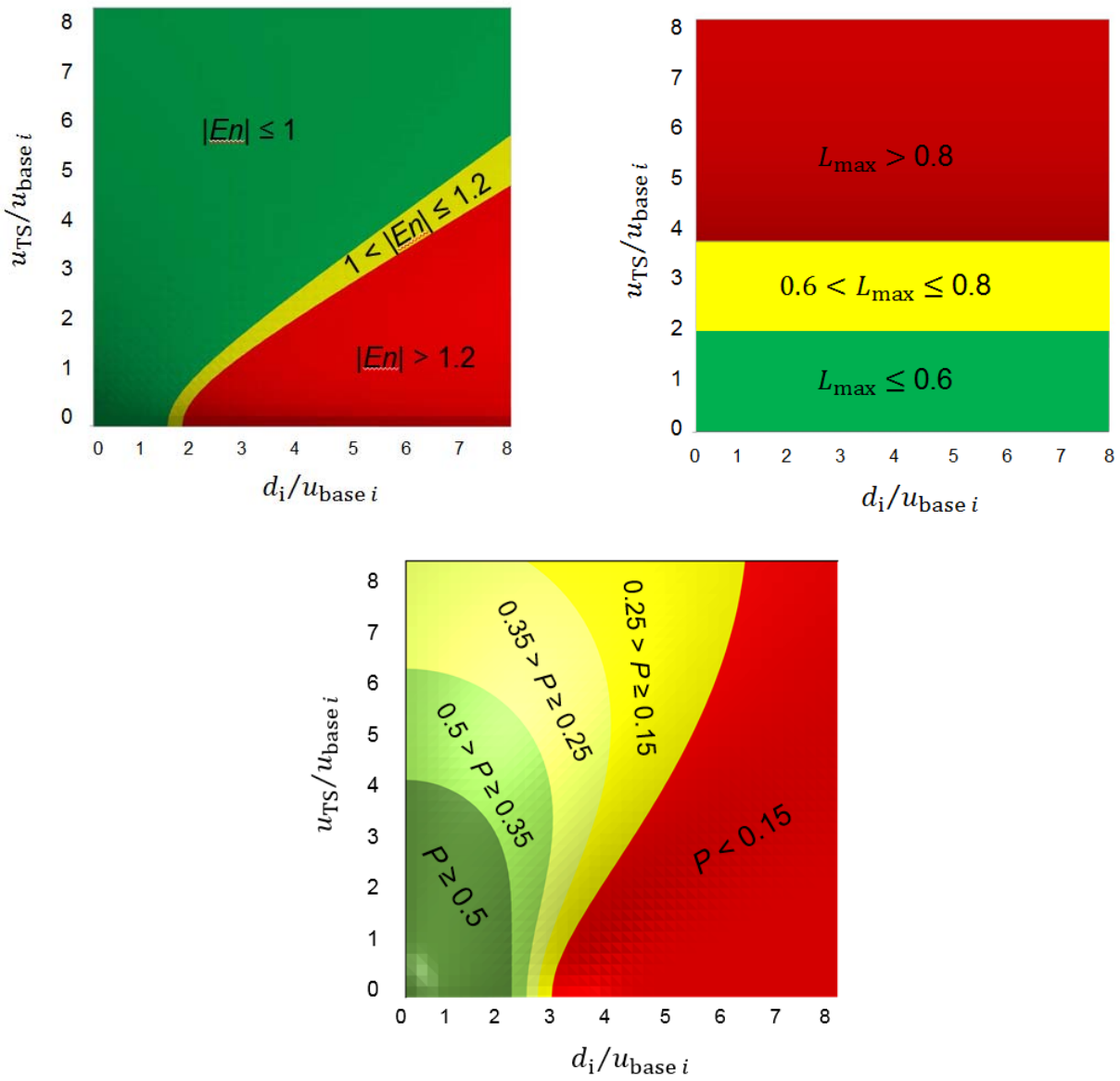


Figure 11. Contour plots of **a)** $|En_i|$ **b)** $L_{\max,i}$ and **c)** P_i for $d_i/u_{base\ i}$ and $u_{TS}/u_{base\ i}$ ranging from 0 to 8.

In the remainder of this section, we will compare our visual assessment of the bi-lateral examples (summarized in Figure 12a) with the four proposed comparison criteria. Figure 12b shows the results for Criterion A over the $d_i/u_{base\ i}$ and $u_{TS}/u_{base\ i}$ ranging from 0 to 8 parameter space. In this and following figures, green, red, and yellow represent passing, failing, and inconclusive results respectively. As shown in Figure 12b, $|En_i| \leq 1$ alone does not find results with large transfer standard uncertainty inconclusive.

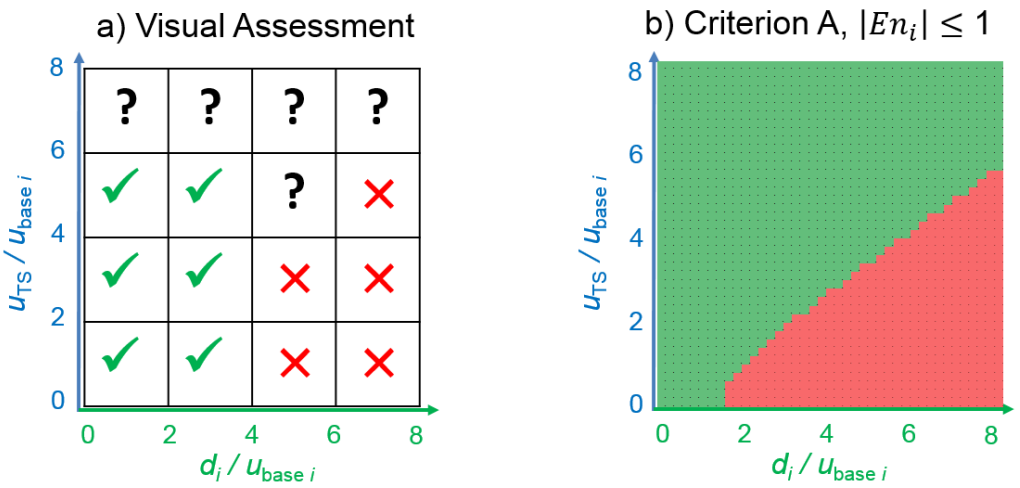


Figure 12. Pass / fail /inconclusive results from our visual assessment and Criterion A for the bi-lateral comparison example.

Figures 13 shows the results for Criteria B and C when applied to the bi-lateral comparison example using various threshold values for R_{th} and L_{th}. These criteria successfully consider cases in the northeast quadrant inconclusive, and do the same for the northwest quadrant, similar to the visual assessment summarized in Figure 12a. For selected threshold values, Criteria B and C will give the same results: Criterion C based on the maximal explanatory power resembles the behavior of the heuristic Criterion B. But note that when specifying a maximum tolerable power loss L_{th} the corresponding upper limit for the uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ depends on u_{TS}, on $u_{\mathrm{base}\,i}$ and on all uncertainties $u_{\mathrm{base}\,j}, j \neq i$ quoted by the remaining laboratories [20].

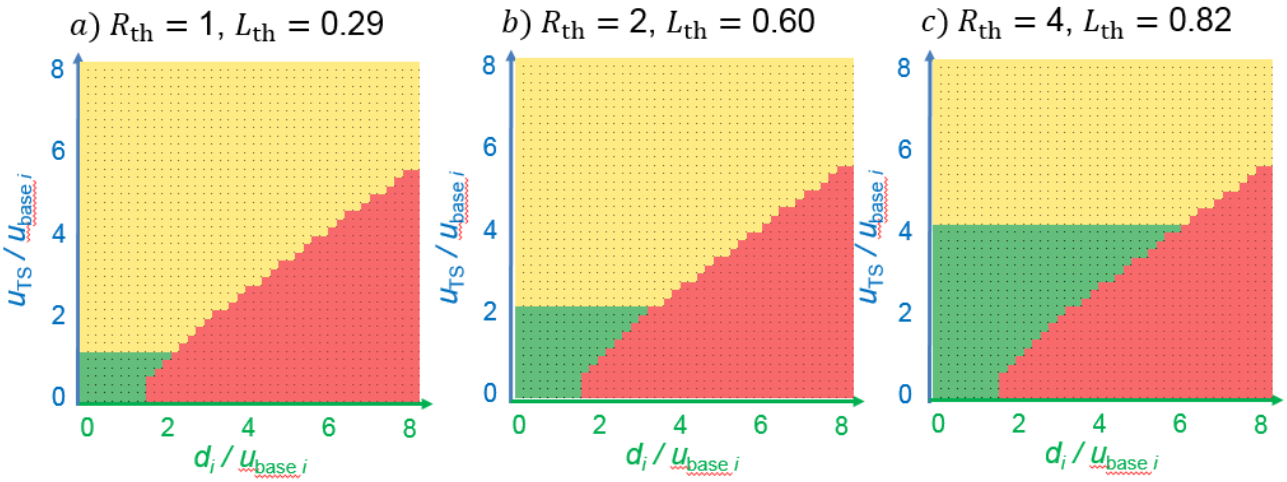


Figure 13. Behavior of Criteria B and C for various threshold values R_{th} and L_{th}.

For Criterion D (Figure 14), the shape of the green passing region expands to the north and northeast as P_{th} is reduced. Criterion D finds comparison results with large $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values inconclusive, even for small $d_{\mathrm{i}}/u_{\mathrm{base}\,i}$ values. Criterion D can be designed to match our visual assessment of the bi-lateral comparison example. In our opinion, a small value of P_{th} is desirable because it does not seem justified to fail participants

that agree well with the CRV in the northwest corner unless the transfer standard uncertainty is quite large: we should consider the possibility that the transfer standard drifted *after* the particular participant gathered their comparison data. However, for $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} > 6$, we recommend that even results with small $d_{\mathrm{i}}/u_{\mathrm{base}\,i}$ be considered inconclusive and this leads us to choose $P_{\mathrm{th}} = 0.35$. Note that the visual assessment has an inconclusive result for $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} = d_{\mathrm{i}}/u_{\mathrm{base}\,i} = 4$, and Criterion D also gives inconclusive results in the central part of the parameter space. Hence Criterion D more closely matches our visual assessment than the other criteria.

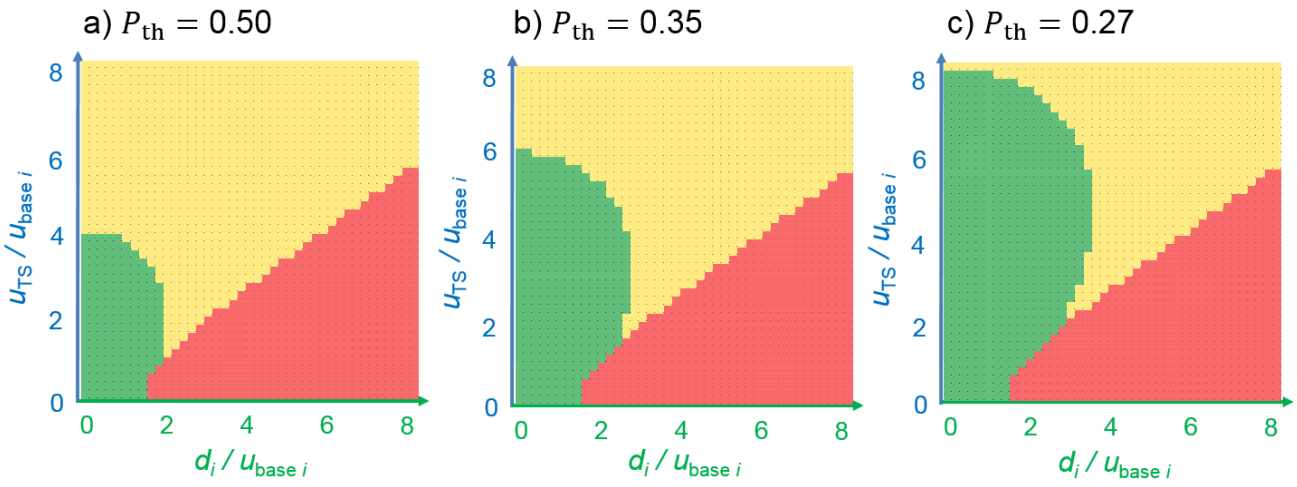


Figure 14. Behavior of Criterion D for various values of P_{th}.

9. Increasing CMCs based on Comparison Results

Before widespread application of the propagation of uncertainties approach in the GUM, many laboratories based their uncertainty statements on the results of comparisons and reproducibility data. At the present time, comparison results are used to validate uncertainty statements, not as an input to calculate them. The WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty [2] state: "There is no established approach for including comparison results in a laboratory's CMCs. A laboratory that obtains unsuccessful comparison results must conduct diagnostic tests and re-examine their uncertainty analysis and revise their CMCs or improve their measurement system."

There may be a clear explanation for a failing result exposed by a root-cause analysis carried out by the participating lab after Draft A of the comparison report is distributed. This effort may result in new values of x_i or $u_{\mathrm{base}\,i}$. The Mutual Recognition Arrangement clearly states that altering these values after other participants' results are revealed is not allowed (except under extenuating circumstances and with the agreement of all participants). But the process of conducting and reporting a comparison and deciding whether CMCs are acceptable are not strictly linked. In fact, the Mutual Recognition Arrangement states that evidence other than comparisons can be used to support CMCs and this is appropriate. A reasonable approach is to report the comparison results without changes in the originally reported data but allow a failing participant the opportunity to explain the cause of the discrepancy, if it is discovered during the period between Draft A and the final version of the comparison report. In this way a comparison report may indicate

that a participant has failed, but that the comparison process has led to improvement and the Pilot, other participants, and the Working Group accept the "failed" lab's CMCs based on their root-cause analysis.

Another circumstance may occur. After the root-cause analysis, no (or insufficient) explanation is found for the discrepant result. In such cases, the question arises: what CMC uncertainty ***is*** supported by the comparison results? In this case, the results of the comparison can be used to recommend the minimum CMC uncertainties that should be accepted by reviewers. For each failing or inconclusive lab result, we can find the smallest additional uncertainty that would need to be added to their $u_{\mathrm{base}\ i}$ to obtain a passing result.

10. Proposed Criteria applied to Real Comparison Data Sets

Next we will apply the commonly used Criterion A ($|En_i| \leq 1$), and the three criteria proposed herein to selected real comparison data sets. We will use $R_{\mathrm{th}} = 2$, $L_{\mathrm{th}} = 0.6$, and $P_{\mathrm{th}} = 0.35$. Note that the CRV values here may not precisely match those in the original comparison reports because some Pilots used uncertainty weighted best-fit curves (a comparison reference curve) [18] while we have processed data at each set point independently from the other set points. Also, Pilot labs used the sample standard deviation s_i to quantify the Type A uncertainty of the participant's reported result and we have used the sample standard deviation of the mean, $s_i/\sqrt{n}$ (as shown in Eqn. 6).

For the bi-lateral examples presented earlier in this paper, we assumed that the Type A component of the comparison uncertainty, $s_i/\sqrt{n}$, was negligible. In our application of the criteria to real data sets, we do not ignore the Type A component $s_i/\sqrt{n}$. It is included in the calculation of u_{x_i} via Eqn. 6 and via u_{d_i}, reduces En_i, making it easier to pass Criterion A and the portions of Criteria B, C, and D that rely on En_i. Larger $s_i/\sqrt{n}$ values also influence x_{CRV} and $u_{x_{\mathrm{CRV}}}$, which in turn influence the values of $L_{\mathrm{max},i}$ and P_i. Larger $s_i/\sqrt{n}$ increases the loss of power $L_{\mathrm{max},i}$ and decreases P_i, and therefore, larger values of $s_i/\sqrt{n}$ reduce the chances of passing results for both Criteria C and D. This is appropriate: the purpose of including $L_{\mathrm{max},i}$ and P_i in the criteria is to identify inconclusive results that are caused by u_{TS} and $s_i/\sqrt{n}$.

Figure 15 shows d_i for all participants at all of the flow set points in a EURAMET low pressure gas flow comparison [12]. Figure 16 shows the results for the 2 m^3/h set point data in the same format described above and shows the pass / fail / inconclusive results when the four criteria are applied. For seven of the twelve participants, separate values of $u_{\mathrm{base}\ i}$ and $s_i/\sqrt{n}$ were not available and $s_i/\sqrt{n}$ was assumed negligible for those labs. The values of $u_{\mathrm{TS}}/u_{\mathrm{base}\ i}$ are less than 0.97 for all participants and there are negligible differences between the $2u_{d_i}$ and $2u_{\mathrm{base}\ i}$ error bars. As we would expect for a comparison where the $u_{\mathrm{TS}}/u_{\mathrm{base}\ i}$ are less than 1, all four criteria deliver the same pass / fail results and there are no inconclusive results. The expanded uncertainty of the CRV is represented in Figure 16 by the shaded region surrounding $d_i = 0$.

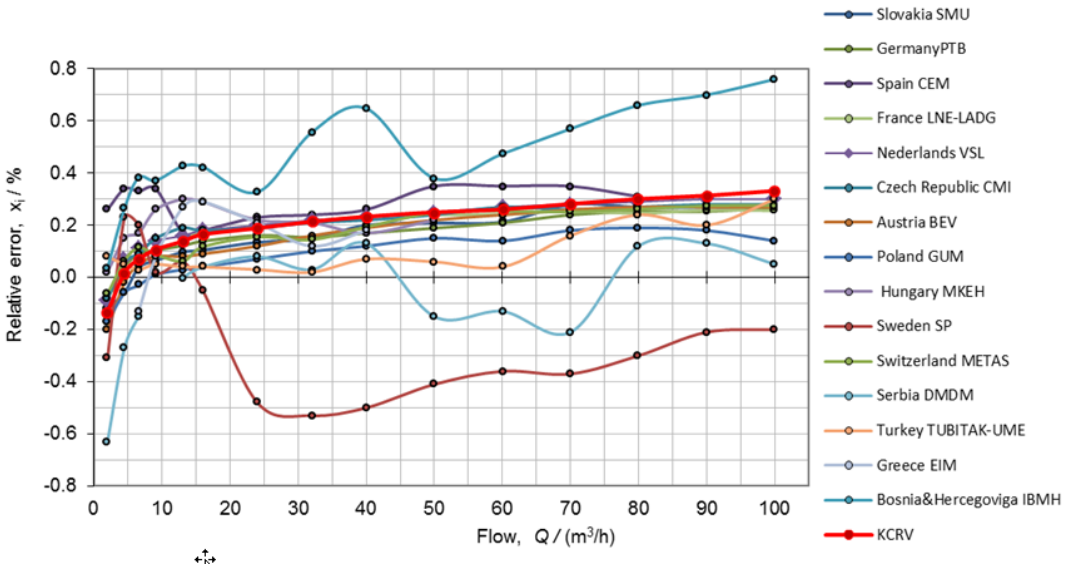


Figure 15. The CRV and reported results for a low-pressure gas flow comparison [12].

	1	2	3	4	5	6	7	8	9	10	11	12
A) $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
B) $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq 2$, $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
C) $L_{\mathrm{max},i} \leq 0.6$, $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
D) $P_i \geq 0.35$, $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓

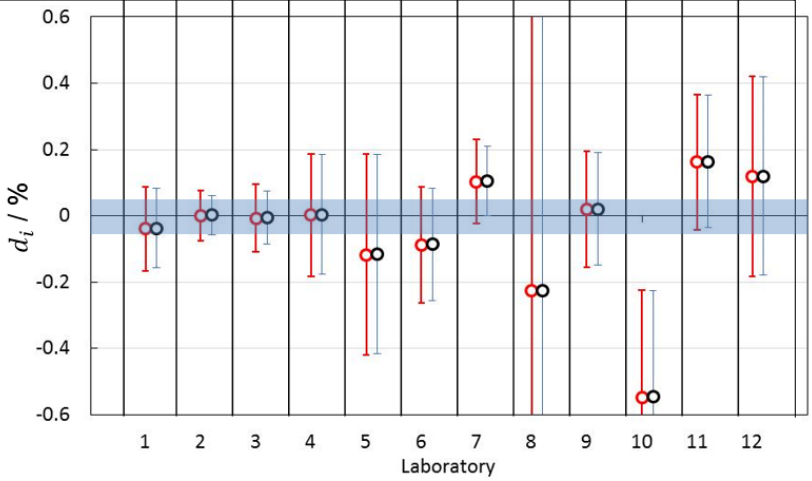


Figure 16. The PDFs, error bars, and criteria results for the 2 m^3/h set point for reference [12]. The shaded region surrounding $d_{\mathrm{i}} = 0$ shows the 95 % confidence uncertainty for the CRV.

The next two data sets were selected because they have large $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values. Figure 17 presents data from the 3.8 L/min set point of the hydrocarbon liquid flow proficiency test described in reference [18] (also shown in Figure 5). For this data set, all $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values are greater than 2 (ranging from 2.2 to 5.7) and therefore, Criteria B and C report all participant's results as inconclusive. Criterion D indicates five labs' results are inconclusive that the $|En_i| \leq 1$ criterion would call equivalent and passes seven labs that Criteria B and C

would deem inconclusive. Criterion D results for Laboratories 2 and 8 are interesting because they fall on either side of the $P_i \geq 0.35$ threshold: Lab 2 passes while Lab 8's results are inconclusive. The $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values for Labs 2 and 8 are 5.7 and 3.4 respectively and on this basis alone, one might expect Lab 8 to receive the more favorable result. But Laboratory 2 has a larger content probability (0.41) than Lab 8 (0.34) because its $u_{\mathrm{base}\,i}$ PDF coincides more closely with that of the CRV (as represented by the shaded region of the plot). Although Lab 3's results look similar to Lab 8's, Lab 3's content probability is larger (0.51).

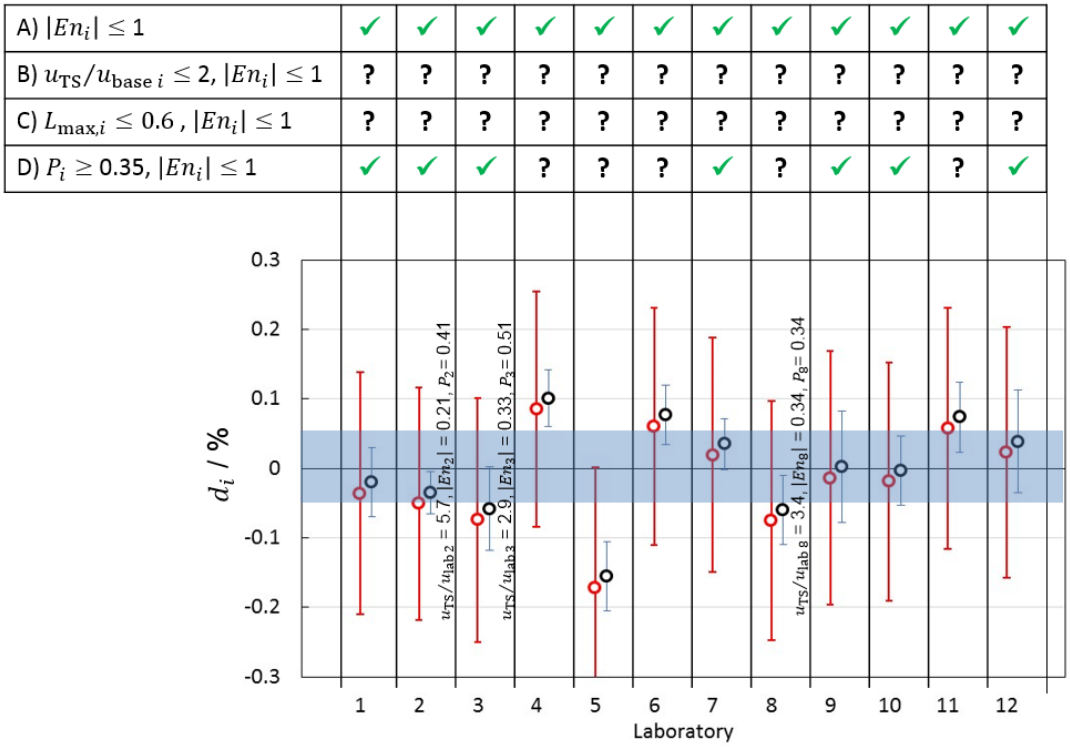


Figure 17. The PDFs, error bars, and criteria results for the 3.8 L/min set point for reference [18].

Figure 18 shows criteria results for the 40 m^3/h set point of a liquid flow comparison that had $u_{\mathrm{TS}}/u_{\mathrm{lab}\,i}$ values between 0.47 and 4.7 [22]. All four criteria found Lab 4's results non-equivalent to the CRV. Three of the labs passed the $|En_i| \leq 1$ criterion but had $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} > 2$ and hence Criteria B and C found those three results inconclusive. Criterion D differs from B and C for three labs and we will examine two of these cases more closely. Criteria B and C find Lab 1's results inconclusive because $u_{\mathrm{TS}}/u_{\mathrm{base}\,1} = 3.6$ and there is significant loss of power. However, Criterion D finds Lab 1's results to be passing because the content probability is 0.68, showing good coincidence of the PDFs for the reference standard and the CRV. The relatively large content probability can be visually assessed by the overlap of the blue error bars representing $2u_{\mathrm{base}\,1}$ with the shaded region representing the expanded uncertainty of the CRV. Turning to Lab 6, $u_{\mathrm{TS}}/u_{\mathrm{base}\,6}$ is 1.8 and $|En_6|$ is 0.97, a passing result for Criteria B and C, but the content probability P_6 is 0.02, less than threshold value used here (0.35), leading to an inconclusive result for Lab 6 from Criterion D. The low content probability can be visually assessed in Figure 18 by the poor overlap of the blue error bars that represent $2u_{\mathrm{base}\,6}$ and the shaded region that represents $2u_{\mathrm{CRV}}$.

A) $\lvert En_i \rvert \leq 1$	✓	✓	✓	×	✓	✓	✓
B) $u_{\mathrm{TS}}/u_{\mathrm{base}\,i} \leq 2$, $\lvert En_i \rvert \leq 1$	?	?	✓	×	✓	✓	?
C) $L_{\mathrm{max},i} \leq 0.6$, $\lvert En_i \rvert \leq 1$	?	?	✓	×	✓	✓	?
D) $P_i \geq 0.35$, $\lvert En_i \rvert \leq 1$	✓	?	✓	×	✓	?	✓

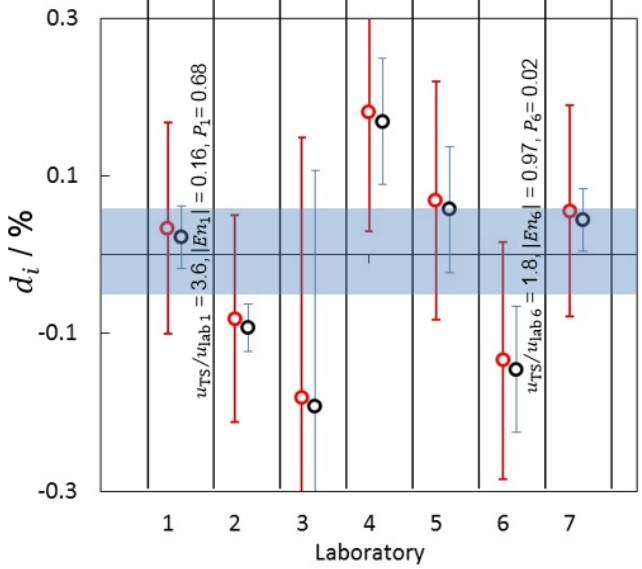


Figure 18. The PDFs, error bars, and criteria results for the 40 m^3/h set point for reference [22].

11. Summary and Conclusions

Realistic assessments of the equivalence of laboratories and CMCs must quantify transfer standard uncertainty and consider u_{TS} effects on comparison results. The bi-lateral comparison example in Figure 6 shows that for large $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values, the commonly used $|En_i| \leq 1$ criterion indicates equivalency when our visual assessment indicates the result should be inconclusive. Criterion A ($|En_i| \leq 1$, current practice) leads to either passing or failing results. This is insufficient in the presence of uncertainty in the transfer standard as demonstrated by several case studies. A third possible outcome needs to be included, namely inconclusive.

We proposed several new pass / fail / inconclusive criteria and studied their behavior when applied to the bi-lateral comparison example and data from three real comparisons. Criterion B is based on a heuristic proposal: limit the maximum acceptable value of $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$. We demonstrated that calculating the maximum loss of power $L_{\mathrm{max},i}$ enables quantification of the deteriorating effect of a transfer standard's instability on the relevance of calculated degrees of equivalence. Criterion C uses the loss of explanatory power of the test $L_{\mathrm{max},i}$ to assess whether results should be considered inconclusive. We proposed to restrict $L_{\mathrm{max},i}$ to be no more than 0.6 for a laboratory to pass a comparison. It happens that Criteria B and C are largely equivalent for appropriate threshold levels.

We defined P_i, the probability that the comparison reference value falls within the 95 % confidence bounds of a participant's results using the base Type B uncertainty of a participant's reference standard. Criterion D

uses $|En_i|$ and P_i to mimic our visual assessment of comparison results. The passing region of Criterion D can be enlarged to include results with larger transfer standard uncertainty by increasing the threshold value P_{th}. We selected $P_{\mathrm{th}} = 0.35$ because it will result in inconclusive results for all $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values greater than 6 as well for some values of $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values less than 6 with intermediate levels of $d_{\mathrm{i}}/u_{\mathrm{base}\,i}$ (see Figure 14b). Note that earlier versions of Criterion D [19] passed labs that agreed with the CRV within $2u_{\mathrm{base}\,i}$ regardless of the value of P_i. But after further consideration, we revised Criterion D because for large $u_{\mathrm{TS}}/u_{\mathrm{base}\,i}$ values, agreement may be due to random transfer standard calibration changes. The present version of Criterion D is simpler and is strongly based on the probability P_i: when the PDF of the comparison reference value is broad due to large transfer standard or Type A uncertainty and the participant's PDF is narrow, P_i will be small.

We note that Criteria C and D are inherently different in regard to their underlying statistical concepts. The explanatory power utilized in Criterion C is based on classical (frequentist) hypothesis testing which makes probability statements in terms of sampling distributions and is thus often applied to evaluate the expected performance of a test. Hence, it is well suited for designing and optimizing key comparison plans, e.g., to assess the impact of transfer standards of different quality on the conclusiveness of a future key comparison. On the other hand, it is also often desirable to be able to use these methods for a retrospective assessment of a completed key comparison, as done here. Criterion D is based on a Bayesian viewpoint by assigning a degree of belief distribution in view of the available data. Once the key comparison results have been compiled, Criterion D can be applied to assess the conclusiveness of that key comparison. For these reasons we think that both criteria provide valuable insights for the assessment of the conclusiveness of a key comparison.

The threshold values L_{th} and P_{th} used in Criteria C and D are arbitrary and different values can be justified. For example, when we are working with $k = 2$ or approximately 95 % confidence intervals, should there be a "warning level" to account for the non-zero probability of a reported value beyond $2u_{d_i}$? Furthermore, it is important to find values that are appropriate for the particular measurand. Future work will apply Criteria C and D to more comparisons and assess appropriate threshold values L_{th} and P_{th} for various measurands.

The Type A component $s/\sqrt{n}$ is included in all four comparison criteria because it affects En_i, x_{CRV}, $u_{x_{\mathrm{CRV}}}$, $L_{\mathrm{max},i}$, and P_i. Colleagues inform us that for some measurands, $s/\sqrt{n}$ is the largest contributor to CMC uncertainty. However, $s/\sqrt{n}$ was not a dominant component in the examples we have used here. Therefore situations where $s/\sqrt{n}$ is large deserve future consideration. In the meantime, we can make a few observations. For some measurands (e.g. flow or pressure), two devices can be calibrated simultaneously and a correlation analysis [23] applied to the results to quantify what portion of the Type A uncertainty is due to the device under test and what portion is due to the laboratory's reference standard. This is rarely done because of the extra work required and $s/\sqrt{n}$ is generally small relative to u_{base}. Without the correlation analysis, $s/\sqrt{n}$ behaves much like u_{TS} in that they both broaden the PDF of the reported value and obscure the conclusiveness of a comparison. Large $s/\sqrt{n}$ may indicate unknown or underestimated Type B components in the reference standard [11], and in these cases, it is important that a Type A component is included in the CMC uncertainty calculation.

Acknowledgements: The authors acknowledge Will Guthrie and Antonio Possolo of the NIST Statistical Engineering Division for their input. We also acknowledge the members of the Working Group for Fluid Flow, the members of the WGFF Pass / Fail Criteria sub-Group, and particularly Gudrun Wendt , Miroslava Benkova, Takashi Shimada, and Yoshiya Terao.

Nomenclature

d_i	Degree of equivalence $= x_i - x_{\mathrm{CRV}}$.
δ_i	Quantity, or measurand, being estimated by d_i.
a_i, b_i	The 2.5$^{\mathrm{th}}$ and 97.5$^{\mathrm{th}}$ percentile confidence limits for lab i based on $u_{\mathrm{base}\ i}$.
En_i	Standardized degree of equivalence between a lab i and the key comparison reference value, $= d_i/2u_{d_i}$.
ε	Difference between the transfer standard and reference flow measurements.
i	Participating lab index.
k	Coverage factor associated with a specified confidence level.
$L_{\mathrm{max},i}$	Maximum loss of power due to transfer standard instability for participant i.
L_{th}	Threshold value of maximum loss of power used in comparison Criterion C.
$\mathrm{N}(\mu, \sigma)$	Normal, Gaussian probability distribution with mean μ and standard deviation σ.
n	Number of measurements made at a set point.
P_i	Probability content of the intervals (a_i, b_i) under the comparison reference value (CRV) distribution.
P_{th}	Threshold probability used in comparison Criterion D.
R_{th}	Threshold value of $u_{\mathrm{TS}}/u_{\mathrm{base}\ i}$ used in comparison Criterion B.
s	The standard deviation of a set of measurements, sample standard deviation.
T	Temperature
$u_{\mathrm{CMC}\ i}$	Standard uncertainty for a lab's calibration and measurement capabilities (CMCs).
u_{drift}	Long term reproducibility (calibration drift) of the transfer standard.
$u_{\mathrm{base}\ i}$	Type B standard uncertainty of the participating laboratory's reference standard obtained by using the law of propagation of uncertainty as described in the GUM [8].
u_T, u_P, u_{prop}	Standard uncertainties due to temperature, pressure, and property sensitivities of the transfer standard.
u_{TS}	Standard uncertainty of the transfer standard, accounting for uncertainty due to transfer standard drift during the comparison, temperature sensitivities, pressure sensitivities, property sensitivities, etc.
u_{x_i}	Standard uncertainty of the reported value from the participating laboratory, accounting for uncertainty due to base reference standard uncertainty, transfer standard uncertainty, and standard deviation of the mean of *n* measurements at each set point.
$u_{x\ \mathrm{CRV}}$	Standard uncertainty of the comparison reference value (CRV).
u_{d_i}	Standard uncertainty of the difference between a participant's reported result and the CRV.
x_i	Reported value of the measurand by the participating laboratory i.
x_{CRV}	The comparison reference value.

[1] *The CIPM Mutual Recognition Arrangement*, 2003, http://www.bipm.org/en/cipm-mra/
[2] The BIPM Key Comparison Database, http://kcdb.bipm.org.
[3] Cox M. G., *Evaluation of Key Comparison Data*, Metrologia, **39**, 589 to 595, 2002.
[4] Toman, B., and Possolo, A., *Laboratory Effects Models for Interlaboratory Comparisons*, Accred. Qual. Assur., **14**(10), 553 to 563, 2009.
[5] Elster, C., and Toman, B., *Analysis of Key Comparison Data: Critical Assessment of Elements of Current Practice with Suggested Improvements*. Metrologia, **50**(5), 549, 2013.
[6] Elster, C., and Toman, B., *Analysis of Key Comparisons: Estimating Laboratories' Biases by a Fixed Effects Model Using Bayesian Model Averaging*. Metrologia, **47**(3), 113, 2010.
[7] *WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty*, Working Group for Fluid Flow, October 21, 2013, http://www.bipm.org/utils/en/pdf/ccm-wgff-guidelines.pdf.
[8] *Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008, http://www.bipm.org/en/publications/guides/.
[9] *Calibration and Measurement Capabilities in the Context of the CIPM MRA*, CIPM MRA-D-04, Version 2, September 2010.
[10] *ILAC Policy for Uncertainty in Calibration*, ILAC-P14:12/2010.
[11] Thompson, M. and Ellison, S. L. R., *Dark Uncertainty*, Accred. Qual. Assur., **16**, 483 to 487, 2011.
[12] Benkova, M., Makovnik, S., Mickan, B*., Comparison of the Primary (National) Standards of Low-Pressure Gas Flow*, EURAMET Project No. 1180, EURAMET.M.FF-K6, October, 2014, http://kcdb.bipm.org.
[13]Toman, B. and Possolo, A., *Model Based Uncertainty Analysis in Inter-Laboratory Comparisons*, Conference on Advanced Mathematical and Computational Tools in Metrology and Testing, Paris, France, June 23 to 25, 2008.
[14] Rukhin, A. L., *Weighted Means Statistics in Interlaboratory Studies*, Metrologia, **46**, 323 to 331, 2009.
[15] Hartig, F., Bosse, H., and Krystek, M., *Recommendations for Unified Rules for Key Comparison Evaluation*, Key Engineering Materials, 6 13, pp. 26 to 33, 2014.
[16] Wright, J., Mickan, B., Paton, R., Park, K.-A., Nakao, S.-I., Chahine, K., Arias, R., *CIPM Key Comparison for Low-Pressure Gas Flow: CCM.FF-K6*, Metrologia, **44,** 2007**.**
[17] *Measurement Comparisons in the CIPM MRA*, CIPM MRA-D-05, version 1.5, March 2014.
[18] Wright, J., et al., *A Comparison of 12 US Liquid Hydrocarbon Flow Standards and the Transition to Safer Calibration Liquids*, Cal Lab: The International Journal of Metrology, 30 to 38, 2012.
[19] Wright, J. D., Toman, B., Mickan, B., Wübbeler, G., Bodnar, O., and Elster, C., *Pass / Fail / Inconclusive Criteria for Inter-Laboratory Comparisons*, Proceedings of the 9th International Symposium on Fluid Flow Measurement, Arlington, Va, USA, April 14 to 17, 2015.
[20]Wübbeler, G., Bodnar, O., Mickan, B., and Elster, C., *Explanatory Power of Degrees of Equivalence in the Presence of a Random Instability of the Common Measurand*, Metrologia, **52**, 400 to 405, 2015.
[21] Toman, B. *Bayesian Approaches to Calculating a Reference Value in Key Comparison Experiments*, *Technometrics*, **49**, 81 to 87, 2007.
[22] Wendt, G. and Marfenko, I., *Draft Report on Supplementary Comparison of National Standards for Liquid Flow*, COOMET.M.FF-S2,COOMET Project 406/UA/07, Braunschweig, October 2012.
[23] Mattingly, G. E., *Flow Standards* in *Flow Measurement: Practical Guides for Measurement and Control*, 2nd edition, D. W. Spitzer editor, The Instrumentation, Systems, and Automation Society, Research Triangle Park, North Carolina, 761 to 777, 2001.
[23] *An Introduction to the "Guide to the Expression of Uncertainty in Measurement"*, JCGM 104:2009, http://www.bipm.org/en/publications/guides/.

[24] Toman, B. *Statistical Interpretation of Key Comparison Degrees of Equivalence Based on Distributions of Belief*, Metrologia **44**(2), 14-17, 2007.