Pass / Fail / Inconclusive Criteria for 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

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 2.5th and 97.5th percentile confidence limits for lab i based on $u_{\mathrm{lab}\ i}$.
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.
$L_{P,i}$	Maximal loss of power due to transfer standard instability for participant i.
n	Number of measurements made at a set point.
P_i	Probability coverage of the intervals (a_i, b_i) under the comparison reference value (CRV) distribution.
P_{th}	Threshold probability used in comparison criteria.
s	The standard deviation of a set of measurements, sample standard deviation.
T	Temperature
$u_{\mathrm{CMC}\ i}$, $U_{\mathrm{CMC}\ i}$	Standard uncertainty and expanded uncertainty (95 % confidence level) in a lab's calibration and measurement capabilities (CMCs).
u_{drift}	Long term reproducibility (calibration drift) of the transfer standard.
$u_{\mathrm{lab}\ i}$	Standard uncertainty of the participating laboratory's reference standard obtained by using the law of propagation of uncertainty as described in the GUM [1]. Also called $u_{\mathrm{base}\ i}$ in reference [2].
$u_{\mathrm{repeat,BED}}$	Standard deviation of the mean $s/\sqrt{n}$ of n repeated measurements performed on the Best Existing Device under test.
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 lab 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.

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 (e.g., flow) 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 $|En_i| \leq 1$ criterion does not always correctly assess whether a participant is working within its uncertainty claims. We show comparison results that demonstrate the problem and discuss the "loss of explanatory power" in terms of statistical hypothesis tests. We propose several criteria for assessing a comparison result as passing, failing, or inconclusive and investigate the behavior of En and alternative comparison measures for a range of $d_{\mathrm{i}}/u_{\mathrm{lab}\,i}$ and $u_{\mathrm{TS}}/u_{\mathrm{lab}\,i}$ values. Two of them (Criteria C and D) 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, 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 [3] and the methodology for conducting and processing a comparison have advanced [4]. 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 and Calibration Report Uncertainty

In 2013, the Working Group for Fluid Flow (WGFF) produced the WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty [2], an attempt to have NMIs use a common approach and terminology in their CMC statements. The Guidelines state that CMC uncertainty (U_{CMC}) is an expanded uncertainty that accounts for the following sources of uncertainty:

1) a base uncertainty[1] of the laboratory's reference standard obtained by using the law of propagation of uncertainty as described in the GUM [1] and 2) a repeatability of n calibration results measured using the Best Existing Device (BED), i.e.,

$$U_{\mathrm{CMC}\,i} = ku_{\mathrm{CMC}\,i} = k\sqrt{u_{\mathrm{lab}\,i}{}^2 + u_{\mathrm{repeat,BED}}{}^2}\,. \qquad (1)$$

[1] Note that the quantity called u_{lab} herein is called u_{base} in reference [2].

$U_{\mathrm{CMC}\,i}$ represents the expanded uncertainty at the 95 % confidence level of the average of n calibration results for the Best Existing Device using laboratory i's reference standard. The quantity $u_{\mathrm{repeat,BED}}$ is the standard deviation of the mean[2] of n repeated measurements performed on the Best Existing Device under test and is included as called for by the CIPM [5] and International Laboratory Accreditation Cooperation [6]. The base uncertainty of the laboratory's reference standard $u_{\mathrm{lab}\,i}$ is a critical input to a laboratory comparison [2]. Note that the base or lab uncertainty is intended to be independent of the particular device being calibrated.

3. Inter-Laboratory Comparisons and Transfer Standard Uncertainty

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 "Criteria 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:

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

The uncertainties in Equation 2 are quantified during preliminary tests organized by the Pilot lab. 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 1. 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 not known until the conclusion of the comparison when repeated calibrations to quantify long-term drift are complete. Often, a rectangular distribution is applied to the range of the calibration changes observed in these Pilot lab calibrations, $(\varepsilon_{\max} - \varepsilon_{\min})$:

$$u_{\mathrm{drift}} = \frac{(\varepsilon_{\max} - \varepsilon_{\min})}{2\sqrt{3}}, \qquad (3)$$

although other approaches (such as taking the standard deviation) may also be appropriate, depending on the amount of data available and the drift behavior of the transfer standard.

[2] $s/\sqrt{n}$ where s is the sample standard deviation of the n measurements.

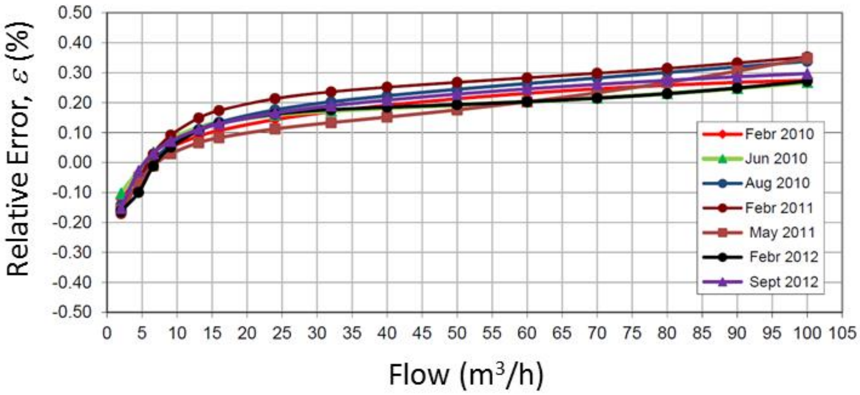


Figure 1. An example of Pilot lab testing of a transfer standard to quantify u_{drift} [7]. In this case, $\varepsilon_{\max} - \varepsilon_{\min} = 0.10$ %, leading to $u_{\mathrm{drift}} = 0.03$ %.

The uncertainty-weighting methods published by Cox [8] are generally used to calculate the CRV, x_{CRV} , its uncertainty, $u_{x_{\mathrm{CRV}}}$, the degrees of equivalence, $d_i = x_i - x_{\mathrm{CRV}}$, and the uncertainty of the degree of equivalence, u_{d_i}. For Cox's approach and independent laboratories (i.e., uncertainties are known perfectly and no covariance between participants), $\frac{1}{{u_{x_{\mathrm{CRV}}}}^2} = \frac{1}{u_{x_1}^2} + \frac{1}{u_{x_2}^2} + \cdots + \frac{1}{u_{x_n}^2}$ and $u_{d_i} = \sqrt{u_{x_i}^2 - u_{x_{\mathrm{CRV}}}^2}$.

The uncertainty of the reported value (called u_{x_i} by Cox) is ***not*** simply the uncertainty of the participant's flow reference ($u_{\mathrm{lab}\ i}$): it also includes 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:

$$u_{x_i} = \sqrt{u_{\mathrm{lab}\ i}^2 + u_{\mathrm{TS}}^2 + \frac{s^2}{n}} \quad . \tag{4}$$

The s^2/n term is the variance of the mean of the n measurements made at each flow set point and quantifies the reproducibility of the measurements made in the participant's lab.

Alternatives to Cox's methods for calculating the CRV and its uncertainty are available [9, 10] and they should be used too. In some flow comparisons, multiple methods have been applied and presented in the comparison reports [11] and the relatively small differences between the CRVs and their uncertainty have 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 Criteria

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,

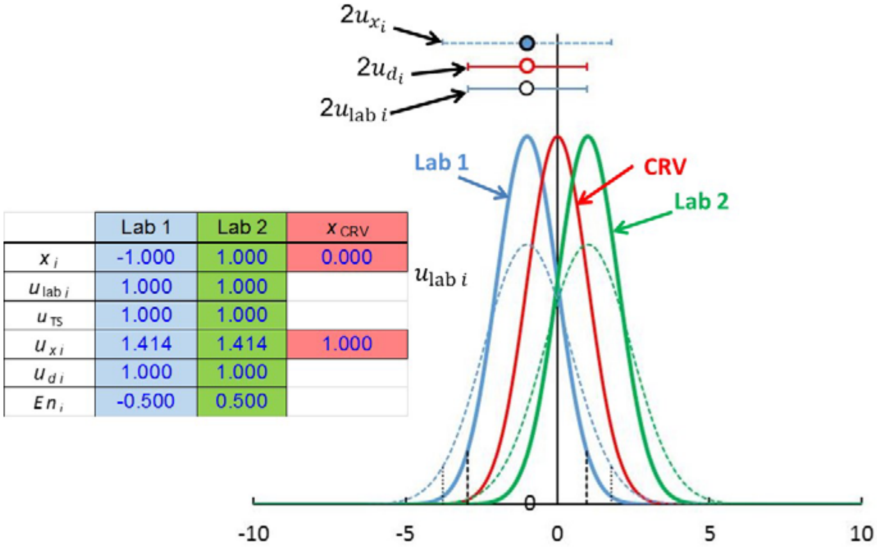
$$En_i = \frac{d_i}{2u_{d_i}} \qquad (5)$$

and what we will call:

Criteria A: Participant i passes if $|En_i| \leq 1$ and fails if $|En_i| > 1$.

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 $u_{\mathrm{lab}\ i}$ 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 hence small En_i. Some graphical examples from a fictitious "bi-lateral comparison example" help to explain.



	Lab 1	Lab 2	x_{CRV}
x_i	-1.000	1.000	0.000
$u_{\mathrm{lab}\ i}$	1.000	1.000	
u_{TS}	1.000	1.000	
$u_{x\ i}$	1.414	1.414	1.000
$u_{d\ i}$	1.000	1.000	
En_i	-0.500	0.500	

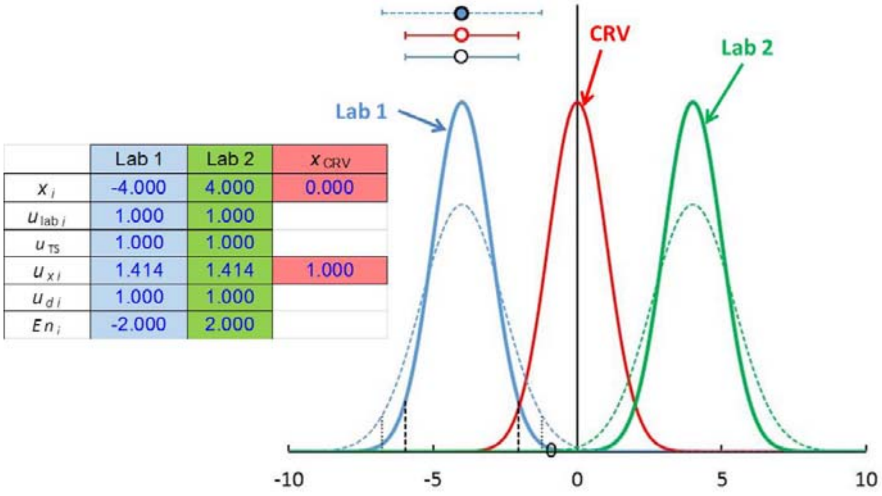
Figure 2. Clearly equivalent: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 1$, $|En_1| = |En_2| = 0.5$, and the participating labs are equivalent. The dashed curves represent the probability density functions for u_{x_i}, i.e. including the transfer standard uncertainty.

In the first bi-lateral comparison example (Figure 2), 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., $u_{\mathrm{lab}\ 1} = u_{\mathrm{lab}\ 2} = 1$, 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 $u_{\mathrm{lab}\ i}$, 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 2 tabulates the quantities for this example and uses a format to present the comparison results that we will use throughout this paper. Figure 2 plots Gaussian probability density functions (PDFs) of the participants' reported values and the CRV. Two versions of the participants' PDFs are shown, one for the lab's base uncertainty $u_{\mathrm{lab}\ i}$ (solid lines), and a second that uses the uncertainty of the reported value u_{x_i} (dashed lines). The high degree of overlap of all the PDFs in Figure 2 is a strong indication of equivalence between the two labs and the CRV.

We can use the mean and the expanded uncertainty of a participating lab ($2u_{\mathrm{lab}\ i}$) to calculate a 95 % uncertainty interval (a_i, b_i) for their measurement where a_i is the 2.5th percentile of the distribution and b_i is the 97.5th percentile. This 95 % confidence interval is represented by dashed, vertical lines and for lab 1 in Figure 2. Three circular symbols represent $x_1$and they have error bars representing the 95 % uncertainty interval based on u_{x_i} (dashed blue), u_{d_i} (red), and $u_{\mathrm{lab}\ i}$ (blue).

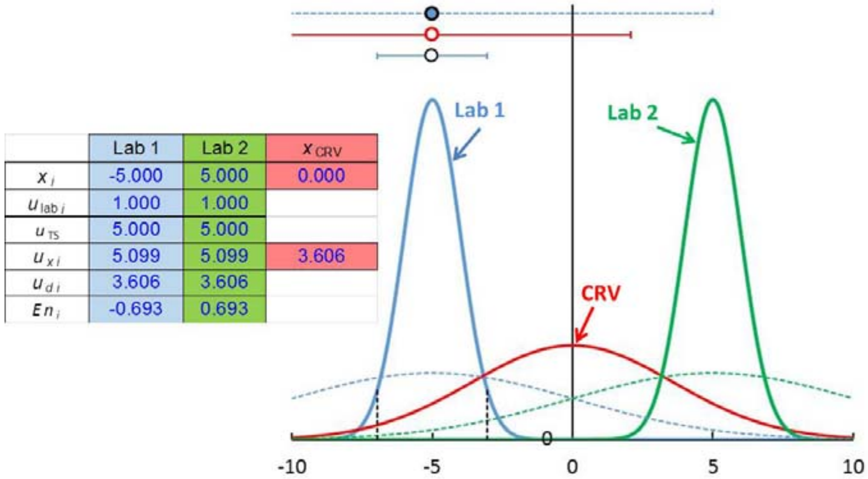
In a second example (Figure 3), the values reported by the two labs are quite different from each other and with the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 1$, $|En_1| = |En_2| = 2$, i.e. by the $|En_i| \leq 1$ criterion, the two labs' results are not equivalent. The lack of overlap of the Lab 1 and Lab 2 PDFs and the various error bars with the CRV at $x = 0$ also indicates that the results are not equivalent.



	Lab 1	Lab 2	x_{CRV}
x_i	-4.000	4.000	0.000
$u_{\mathrm{lab}\ i}$	1.000	1.000	
u_{TS}	1.000	1.000	
$u_{x\ i}$	1.414	1.414	1.000
$u_{d\ i}$	1.000	1.000	
En_i	-2.000	2.000	

Figure 3. Clearly not equivalent: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 1$, $|En_1| = |En_2| = 2$, and the participating labs are not equivalent.

In our third example (Figure 4), the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 5$, 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 $|En_i| \leq 1$ criteria indicates equivalence, in this case a "false positive" result considering the lack of overlap of the two labs' PDFs ($P(|x_1 - x_2| < 3) \approx 0$).



	Lab 1	Lab 2	x_{CRV}
x_i	-5.000	5.000	0.000
$u_{\mathrm{lab}\,i}$	1.000	1.000	
u_{TS}	5.000	5.000	
$u_{x\,i}$	5.099	5.099	3.606
$u_{d\,i}$	3.606	3.606	
En_i	-0.693	0.693	

Figure 4. Inconclusive: the comparison uncertainty ratio $u_{\mathrm{TS}}/u_{\mathrm{lab}\,i} = 5$, $|En_1| = |En_2| = 0.69$ (≤ 1), but the participating labs reported results do not appear to be equivalent.

Our review of past key and regional comparison reports shows that when values for u_{TS} are given, $u_{\mathrm{TS}}/u_{\mathrm{lab}\,i}$ is often larger than 1, and in some cases greater than 5. The example in Figure 4 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 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 $u_{\mathrm{lab}\,i}$.

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 [12] 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 [13] in which $u_{\mathrm{TS}}/u_{\mathrm{lab}\,i}$ ranged between 2.2 and 5.7. Three versions of the results are shown for each lab, 1) the filled circular symbols include the transfer standard uncertainty in the CRV calculation process and the coverage intervals are equal to $2u_{x_i}$ 2) the red open symbols have $2u_{d_i}$ coverage intervals, and 3) the blue open symbols have coverage intervals equal to $2u_{\mathrm{lab}\,i}$. The $2u_{\mathrm{lab}\,i}$ and $2u_{x_i}$ coverages are shown to illustrate the influence of u_{TS} on the results from Criteria A. For laboratories 2, 3, 4, 6, 8 and 11, including the large transfer standard uncertainty 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 $|En_i| \leq 1$ criteria cross the CRV value (0) while the solid blue error bars do not.

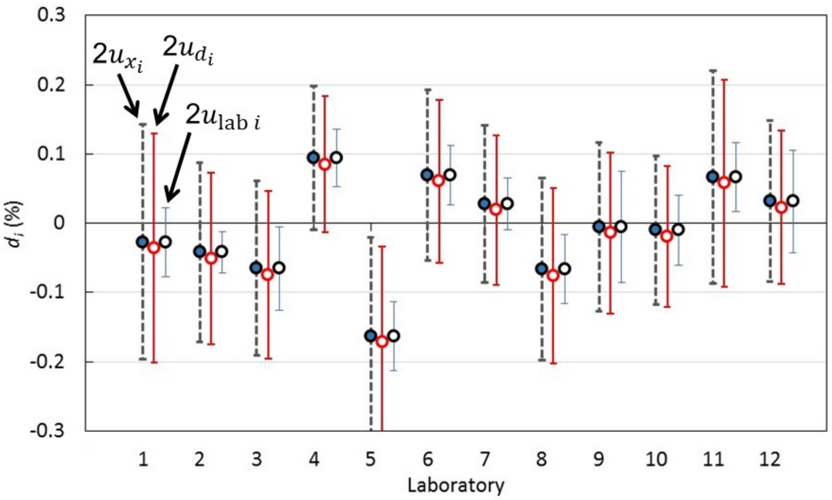


Figure 5. Liquid flow comparison results [13] with three versions of error bars 1) $2u_{x_i}$ (dashed), 2) $2u_{d_i}$ (red), and 3) $2u_{\text{lab}\ i}$ (blue).

5. Behavior of PDFs and *En* over the $d_{\text{i}}/u_{\text{lab}\ i}$ and $u_{\text{TS}}/u_{\text{lab}\ i}$ Parameter Space

We return to the bi-lateral comparison example described in section 4, *i.e.* $x_1 = -x_2$, $u_{\text{lab 1}} = u_{\text{lab 2}}$, and negligible irreproducibility. Figure 6 plots PDFs for $d_{\text{i}}/u_{\text{lab}\ i}$ and $u_{\text{TS}}/u_{\text{lab}\ i}$ 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 blue), $2u_{d_i}$ (red), and $2u_{\text{lab}\ i}$ (blue).

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 with the following criteria. If the reported value agrees with the CRV within the participant's $2u_{\text{lab}\ i}$ claim (i.e. the solid blue error bars cross 0) the reported value is considered equivalent to the CRV (✓). If $|En_i| > 1$, the reported value is considered not equivalent (×). Neither the $2u_{\text{lab}\ i}$ or $2u_{d_i}$ error bars cover the CRV for the cases marked ×. The cases labelled **?** are considered inconclusive because the $2u_{\text{lab}\ i}$ error bars do not cover the CRV, but the $2u_{d_i}$ do (because of transfer standard uncertainty).

The southwest quadrant 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 of Figure 6 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.

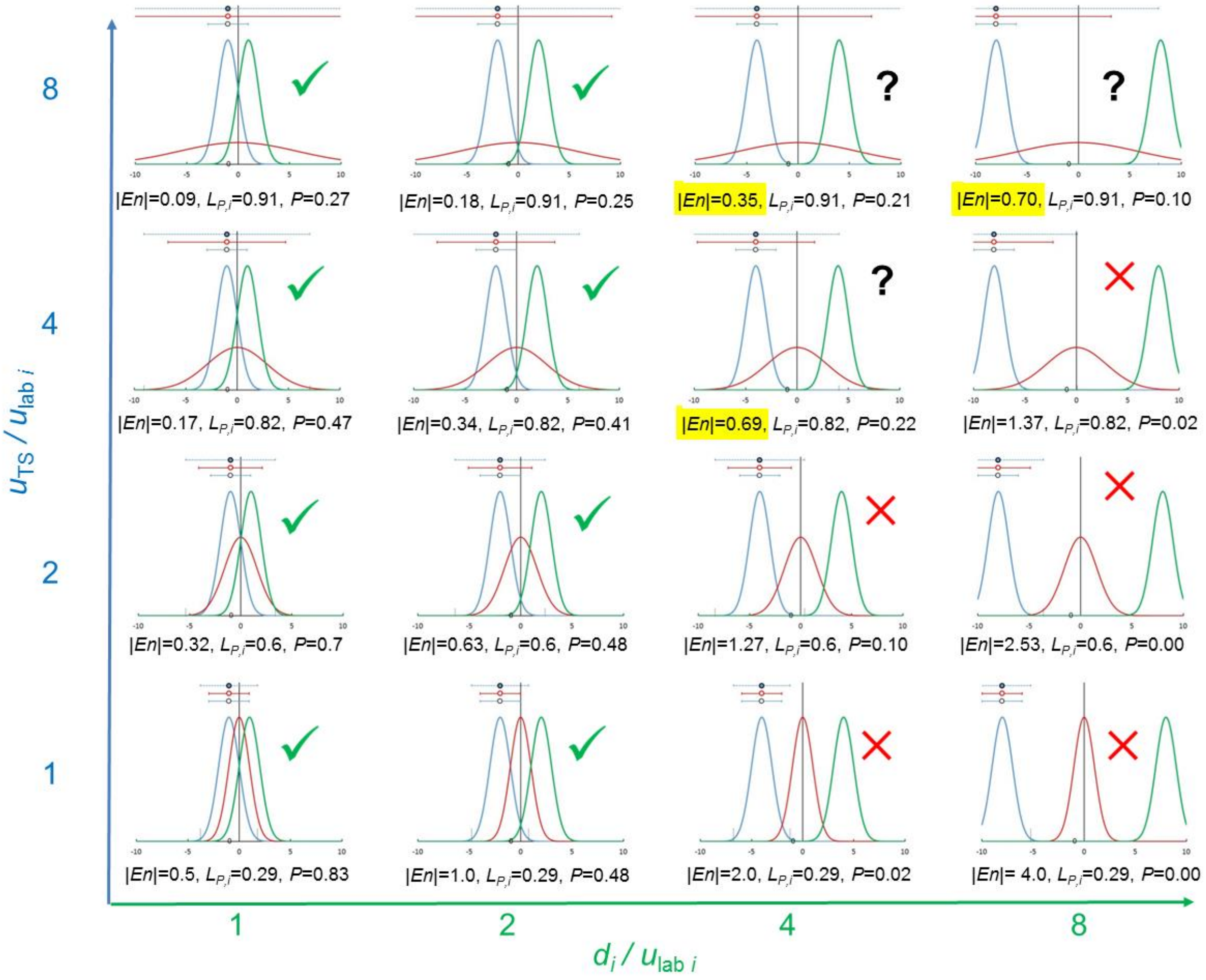


Figure 6. Probability density functions for the bi-lateral comparison example plotted for $d_\mathrm{i}/u_{\mathrm{lab}\ i}$ and $u_\mathrm{TS}/u_{\mathrm{lab}\ i}$ ranging from 1 to 8 with a visual assessment for ✓ (equivalent or passing), × (not equivalent or failing), or **?** (inconclusive).

The northwest quadrant of Figure 6 is deemed passing despite large u_TS, u_{x_i}, and u_{d_i}. One could argue that the cases in this quadrant might only show small d_i because of random transfer standard uncertainty and should be considered inconclusive. But it does not seem appropriate to fail a laboratory when it agrees well with the CRV, even if it may be by chance. Furthermore, if possible we should avoid penalizing a participant in cases where the transfer standard uncertainty is overestimated or the transfer standard may have drifted after a participant completed their calibration of it.

Three of the four cases in the northeast quadrant are shaded yellow because they have $|En_i| \leq 1$ (passing), but we would visually assess them as inconclusive. Inspection of the error bars and PDFs shows that $|En_i| \leq 1$ is due to large transfer standard uncertainty (large u_{x_i} and u_{d_i}), not due to good agreement between the participant's measurement and the CRV. The $|En_i| \leq 1$ criterion does not always match our visual assessment for $d_\mathrm{i}/u_{\mathrm{lab}\ i}$ and $u_\mathrm{TS}/u_{\mathrm{lab}\ i}$ > 2 (the northeast quadrant of Figure 6).

A possible solution to the failure of the $|En_i| \leq 1$ criterion for large u_{TS} is to normalize d_{i} with a quantity other than u_{d_i} ($u_{\mathrm{lab}\ i}$ for instance). Unfortunately, the large transfer standard uncertainty introduces the possibility that a laboratory with good reference standards reports a poor result (see the dashed error bars). Hence normalizing d_{i} with $2u_{\mathrm{lab}\ i}$ will sometimes lead to failing results for labs that are operating within their uncertainty claims and indicates a need for more complex criteria.

The WGFF has discussed a possible pass / fail / inconclusive criteria that we will call:

Criteria B: Participant i passes if $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} \leq 2$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant i otherwise.

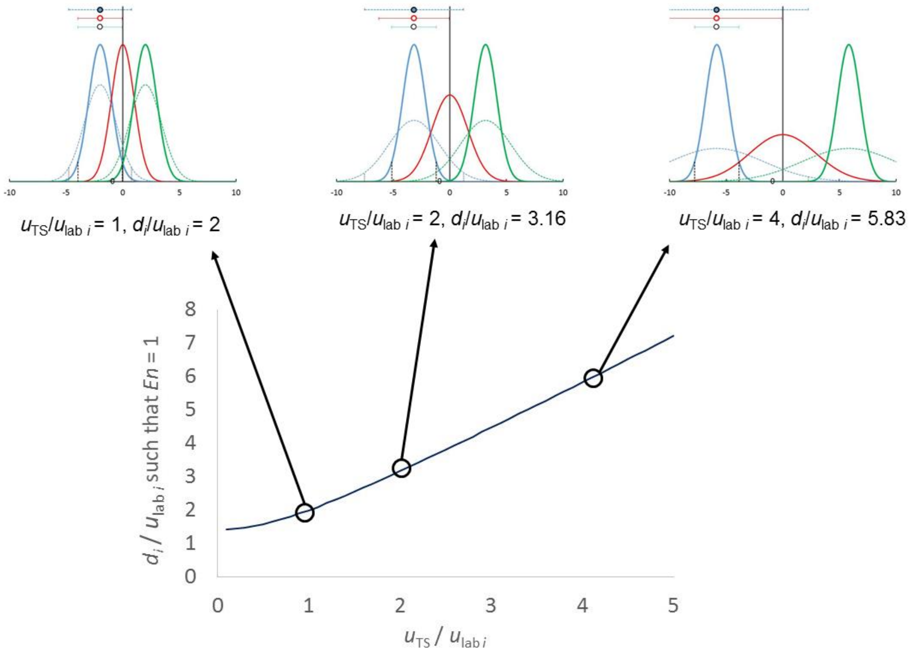


Figure 7. Degree of equivalence that results in $|En_i|$ = 1 versus $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ along with probability density function plots for specific cases of $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 1$, 2, and 4 for the bi-lateral comparison example.

The $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} \leq 2$ criterion avoids participants passing solely because the transfer standard uncertainty was large. The value 2 in the criterion $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} \leq 2$ is subjective. It is based on experience regarding the limitations of transfer standards for some measurands and the fact that the loss of explanatory power is not extreme (see following section). Figure 7 shows that as $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ increases (and explanatory power decreases), the $|En_i| \leq 1$ criterion alone passes participants with large differences from the CRV. For example, if $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 4$, the $|En_i| \leq 1$ criterion alone allows a participant with $d_{\mathrm{i}}/u_{\mathrm{lab}\ i} = 5.83$ to pass. For $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} = 2$, a participant with $d_{\mathrm{i}}/u_{\mathrm{lab}\ i}$ as large as 3.16 passes. 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.

6. Explanatory Power

Wübbeler *et al.* [14] quantitatively assessed the degradation in the "explanatory power" of a comparison due to transfer standard uncertainty. Analytical expressions for the loss of power are given in [13] which can, e.g., be used to analyze the scenarios shown in Figure 6. Figure 8 shows the loss in power for a bi-lateral comparison for a range of $|\delta_1 - \delta_2|$ and the $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ values from Figure 6. The δ_i denote the quantities being estimated by the degrees of equivalence d_{i}. As can be seen the explanatory power decreases monotonically for increasing $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ values.

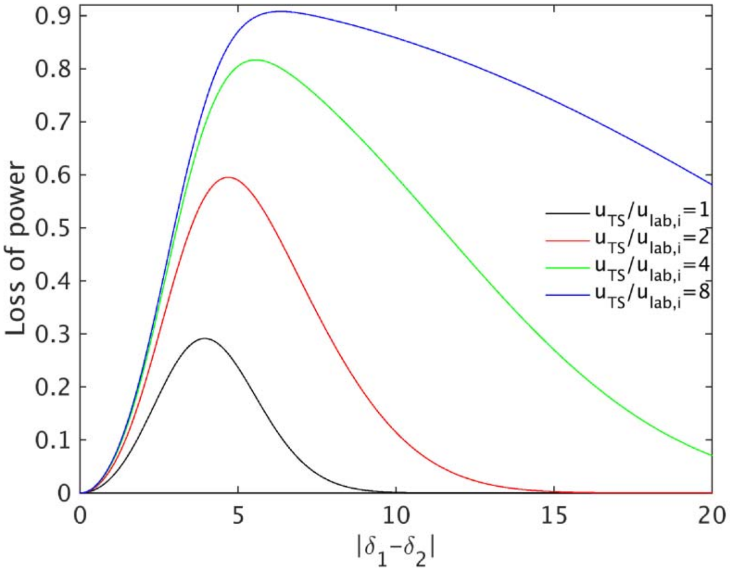


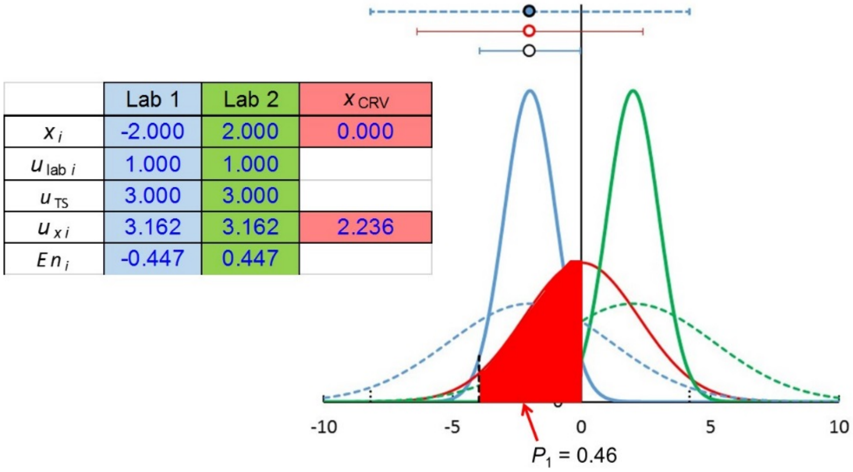
Figure 8. Loss in explanatory power in a bi-lateral comparison as a function of $|\delta_1 - \delta_2|$ for various $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ values where the uncertainties quoted by the two laboratories are assumed to be equal.

The power calculations are useful because they allow us 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 $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$. The power loss can also be utilized to design a reliability criterion for comparison results in the presence of an unstable transfer standard. Setting the maximal tolerable loss of power L_P to a threshold value of $L_{P\ \mathrm{th}}$, e.g., 0.6 results in the following rule:

Criteria C: Participant i passes if the loss of power $L_{P,i}$ satisfies $L_{P,i} \leq 0.6$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant *i* otherwise.

7. Probability Based Criteria

In a comparison, each laboratory reports its own state of knowledge about the measurand which could be represented by a Gaussian probability distribution with mean x_i and standard deviation $u_{\mathrm{lab}\ i}$. Based on this probability distribution the laboratory states a 95% uncertainty interval (a_i, b_i) for the measurand where a_i is the 2.5th percentile of the distribution and b_i is the 97.5th percentile. There is a 0.95 probability that the value of the measurand lies in the interval (a_i, b_i). These percentiles can be easily obtained using various software (for instance by using the function NORM.INV in Excel[3]).



	Lab 1	Lab 2	x_{CRV}
x_i	-2.000	2.000	0.000
$u_{\text{lab}\,i}$	1.000	1.000	
u_{TS}	3.000	3.000	
$u_{x\,i}$	3.162	3.162	2.236
En_i	-0.447	0.447	

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

The CRV incorporates results from n participants and takes into account the uncertainty of the transfer standard. Assuming that the participants provide independent measurements of the same value, the CRV represents the best available estimate of the measurand. At the conclusion of the comparison it is possible to check whether the claims of the individual laboratories were realistic by calculating the probability coverage of the intervals (a_i, b_i) under the distribution based on the CRV, *i.e.* $N\left(x_{\text{CRV}}, u_{x_{\text{CRV}}}\right)$, as represented by the shaded region labelled P_1 in Figure 9. The probability P_i can be calculated in Excel using the following formula: = NORMDIST(NORM.INV(0.975, x_i , $u_{\text{lab}\,i}$), x_{CRV}, $u_{x_{\text{CRV}}}$, TRUE) − NORMDIST(NORM.INV(0.025, x_i, $u_{\text{lab}\,i}$), x_{CRV}, $u_{x_{\text{CRV}}}$, TRUE). When the uncertainty u_{CRV} is small or of similar size to $u_{\text{lab}\,i}$, the coverage probability of (a_i, b_i) should be 0.95 or larger for a well specified initial claim. The probability P_i is listed along with $|En_i|$ in Figure 6. A pass / fail criterion can be set by establishing a threshold value for the probability P_i.

Figure 10 shows contour plots for $|En_i|$, P_i, and $L_{P,i}$ for $d_{\text{i}}/u_{\text{lab}\,i}$ and $u_{\text{TS}}/u_{\text{lab}\,i}$ ranging from 0 to 8. The contour plots use green for possible passing values, yellow for warning levels, and red for values indicating that labs are not equivalent. As described in the discussion of Figure 6, the $|En_i| \leq 1$ criterion is green (passing) in the northeast quadrant, giving false positive results. In contrast, the coverage probability P_i and the loss of power $L_{P,i}$ do not suffer from false positive results for $d_{\text{i}}/u_{\text{lab}\,i}$ and $u_{\text{TS}}/u_{\text{lab}\,i} > 2$ (northeast quadrant). Low coverage 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 $u_{x_{\text{CRV}}}$, (large u_{TS}, northwest corner), or 3) when both of these conditions apply (northeast quadrant).

[3] 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.

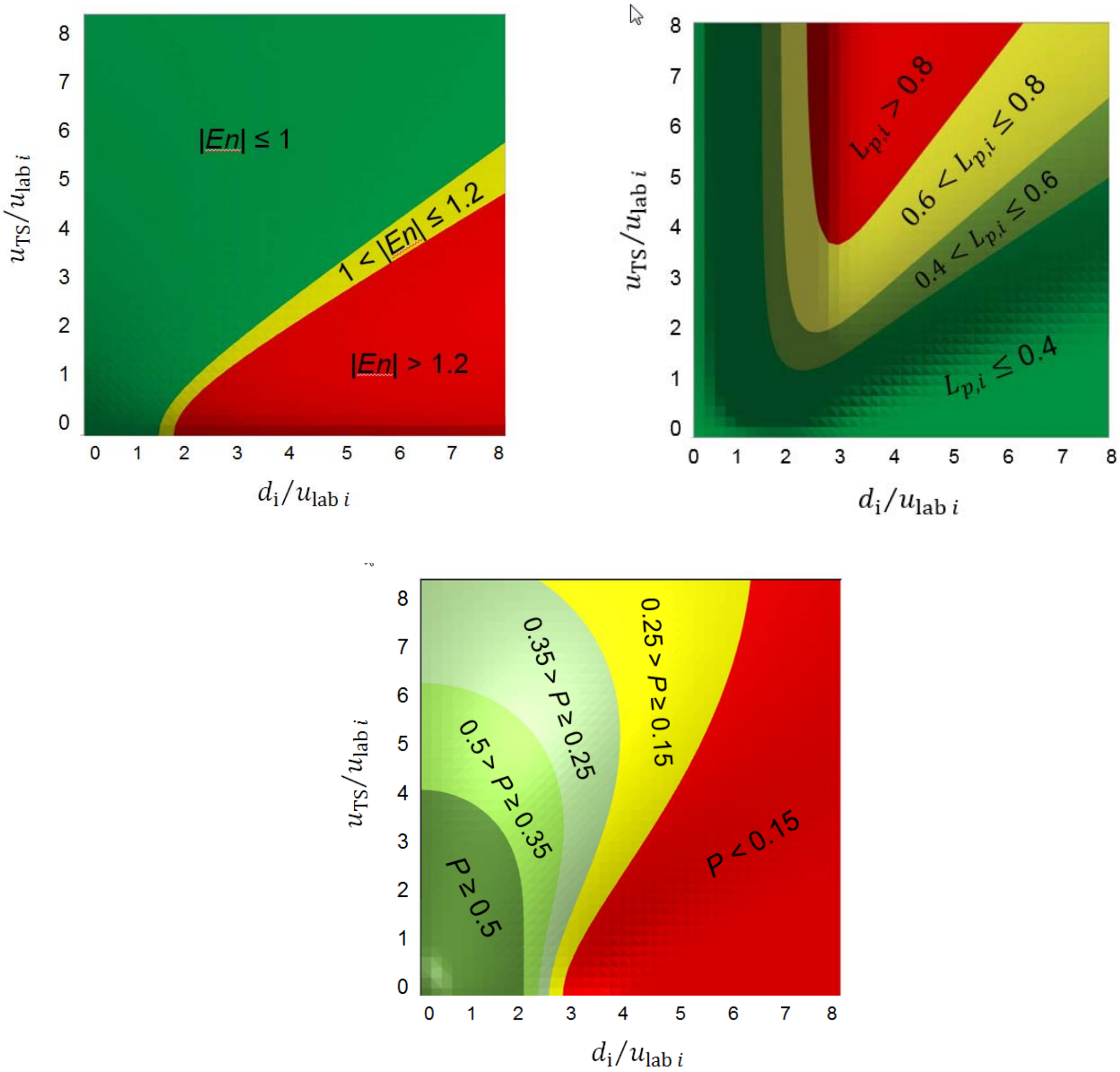


Figure 10. Contour plots of a) $|En_i|$ b) $L_{P,i}$, and c) P_i for $d_\mathrm{i}/u_{\mathrm{lab}\ i}$ and $u_\mathrm{TS}/u_{\mathrm{lab}\ i}$ ranging from 0 to 8.

8. Threshold for Acceptable Probability of Coverage, P_th

Under usual circumstances when most of the laboratories have reasonably specified uncertainties, and u_TS is less than 1, u_{x_CRV} will be smaller than most if not all of the individual $u_{\mathrm{lab}\ i}$. In such circumstances a lab's coverage 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 coverage probability P_i. Clearly 0.95 or higher is best, but lower values could be judged as acceptable. Consider the threshold value $P_\mathrm{th} = 0.5$. The laboratory states that the true value of the measurand lies in the interval (a_i, b_i) with 0.95 probability. If at the conclusion of the comparison the coverage probability of this interval is only 0.5 then there is a probability 0.5 that the measurand is not in this interval at all. So, one might wish to select a higher number than this. On the other hand, the value 0.5 is attractive in that it leads to some very simple rules. Specifically, the coverage probability will always be > 0.5 if $x_{\mathrm{CRV}} \in (a_i, b_i)$ and $u_{\mathrm{CRV}} / u_{\mathrm{lab}\ i} < 1$. So in this case it is not necessary to compute the coverage probability to conclude that the lab's results are acceptable. Conversely, the coverage probability will always be < 0.5 when $x_{\mathrm{CRV}} \notin (a_i,\ b_i)$. It will also be < 0.5 when $x_{\mathrm{CRV}} \in (a_i, b_i)$ but $u_{\mathrm{CRV}} / u_{\mathrm{lab}\ i} > 2.88$. Again, in these cases, conclusion that the lab's results were not acceptable can be drawn without computation of the coverage probability.

Figure 11 plots P_i versus $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ for cases where $|En_i| = 1$ in the bi-lateral comparison example. To remain consistent with the criterion that $|En_i| \leq 1$ while $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} \ \leq 1$, P_{th} = 0.48 and to remain consistent with $|En_i| \leq 1$ while $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i} \ \leq 2$, the P_{th} = 0.22. 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 $P_{\mathrm{th}} < 0.5$. The value of P_{th} used in comparisons may evolve over time as transfer standards improve.

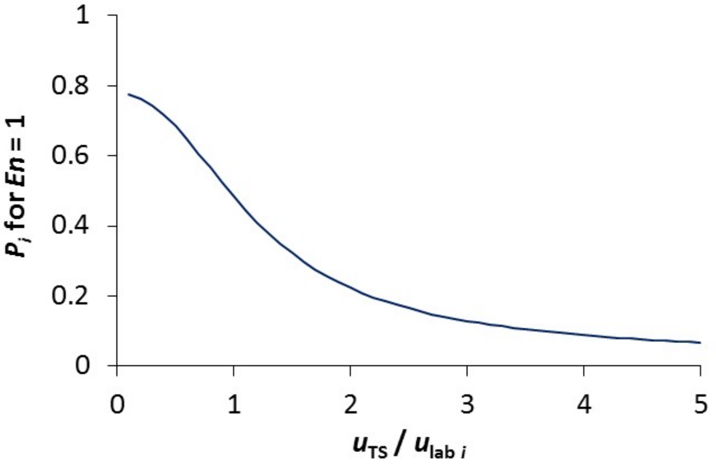


Figure 11. P_i versus $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ for cases where $|En_i| = 1$ in the bi-lateral comparison example.

The standardized degree of equivalence $|En_i|$ and the coverage probability P_i can be used with other success measures to design a set of pass / fail / inconclusive criteria that mimic the visual assessments we made in Figure 6. The examples in the southwest and northwest quadrants of Figure 6 pass due to $|d_i/(2u_{\mathrm{lab}\ i})| \leq 1$. Results in the southeast and northeast quadrants with $|En_i| > 1$ are considered failing. Results with coverage probability P_i below a threshold value P_{th} and where $|d_i/(2u_{\mathrm{lab}\ i})| > 1$ are inconclusive.

<u>Criteria D:</u> Participant i <u>passes</u> if $|d_i/(2u_{\mathrm{lab}\ i})| \leq 1$ or $P_i \geq P_{\mathrm{th}}$, <u>fails</u> if $|En_i| > 1$, and the comparison results are <u>inconclusive</u> for participant i otherwise.

9. Criteria applied to Bi-Lateral Comparison Example

Figure 12 shows the pass / fail / inconclusive results when Criteria A and B are applied to the bi-lateral example for $d_{\mathrm{i}}/u_{\mathrm{lab}\ i}$ and $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ ranging from 0 to 8. Criteria B successfully considers cases in the north-east quadrant inconclusive, but does the same for the northwest quadrant that we visually assessed as passing.

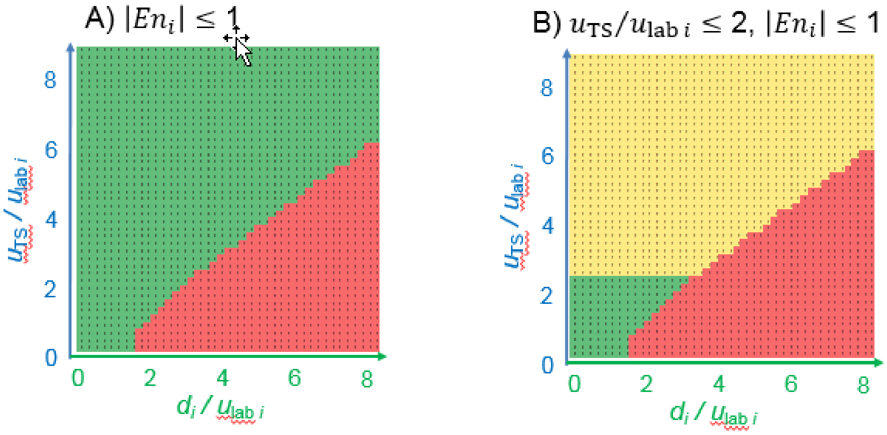


Figure 12. Pass / fail / inconclusive results when Criteria A and B are applied to the bi-lateral comparison example.

Figures 13 and 14 show the results for Criteria C and D applied to the bi-lateral comparison example for various threshold values for $L_{P\ \mathrm{th}}$ and P_{th}. For well-chosen threshold values, both Criteria C and D can match our visual assessment that comparison results in the northeast quadrant are inconclusive. Criteria C has a region of passing results near the diagonal where $d_{\mathrm{i}}/u_{\mathrm{lab}\ i} \approx u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ for $L_{P\ \mathrm{th}}$ greater than 0.45.

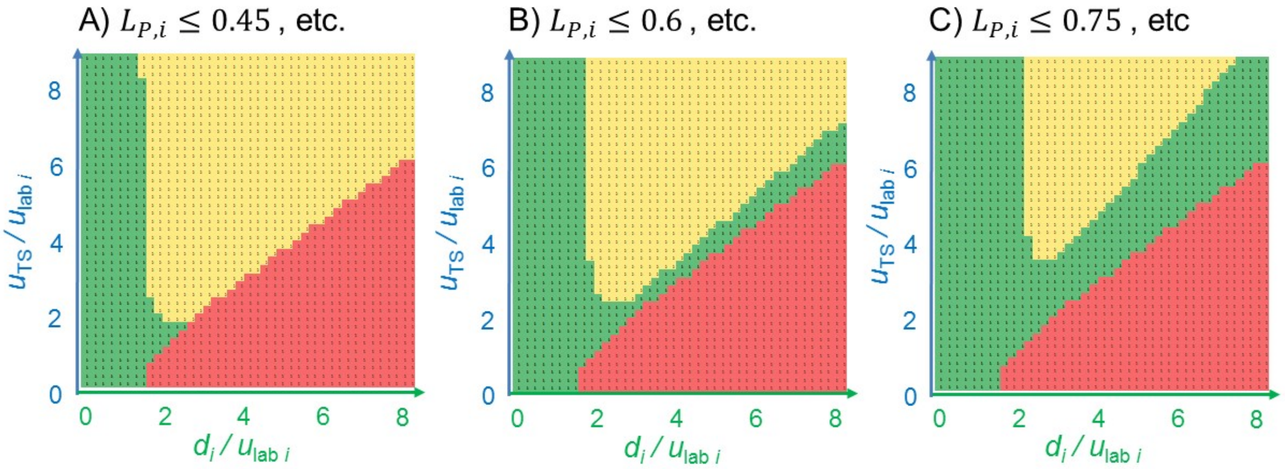


Figure 13. Behavior of Criteria C for various threshold values $L_{P\ \mathrm{th}}$.

For Criteria D with $P_{\mathrm{th}} = 0.44$ or larger, the $|d_i/(2u_{\mathrm{lab}\ i})| \leq 1$ condition determines the entire green passing region in Figure 14a. For $P_{\mathrm{th}} = 0.44$ or larger, the P_i portion of Criteria D is only relevant in determining inconclusive cases. But for P_{th} less than 0.44, the shape of the passing region is expanded to the northeast due to the $P_i > P_{\mathrm{th}}$ portion of the criteria.

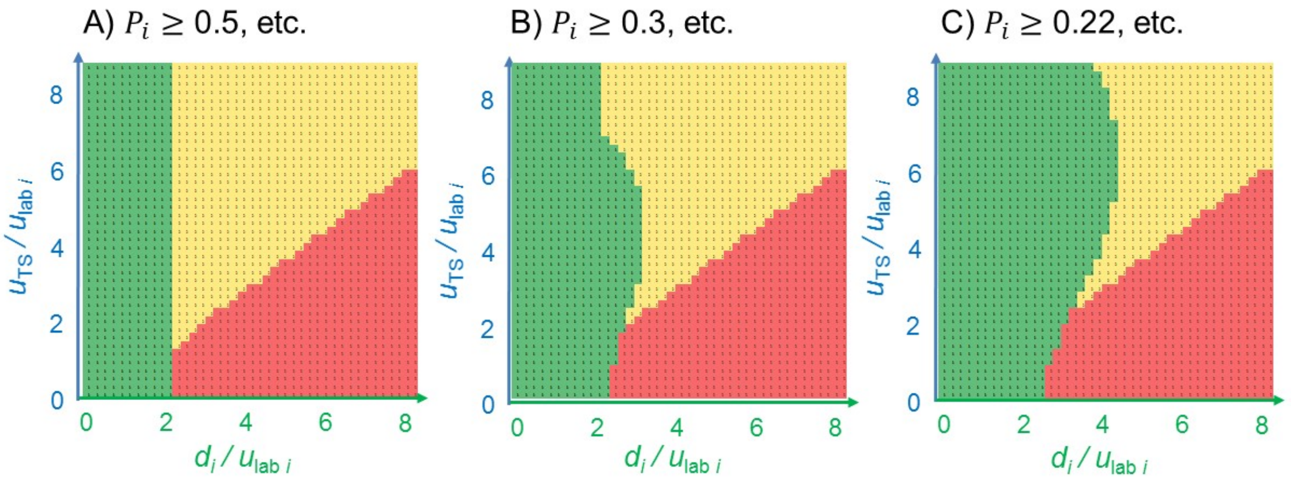


Figure 14. Behavior of Criteria D for various values of P_{th}.

10. Combining Results from Multiple Comparison Set Points

There are other complications to robustly and objectively deciding whether or not a comparison supports a participant's CMC. In many comparisons, more than one set point and transfer standard are tested. For example, in CCM.FF-K6-2006 [11], two sets of 4 transfer standards were calibrated at 7 flow set points to assess a single CMC for each participant. How should these results be combined? (Some Pilots have used the criterion that averaged En values from all set points should be ≤ 1.) When multiple set points of the measurands are used, an arithmetic mean of $|En_i|$ or P_i values is recommended.

11. 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." In such cases, the question arises: what CMC uncertainty ***is*** supported by the comparison results?

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 $u_{\mathrm{lab}\ i}$. 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 MRA states that evidence other than comparisons can be used to support CMCs. 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.

A second circumstance may occur. After the root-cause analysis, no (or insufficient) explanation is found for the discrepant result. 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 $u_{\mathrm{lab}\ i}$ to obtain a passing result.

12. Proposed Criteria applied to Real Comparison Data Sets

Next we will apply the commonly used Criteria A ($|En_i| \leq 1$), and the three criteria proposed herein to selected real comparison data sets. 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) while we have processed data at each set point independently from the other set points.

Figure 15 shows d_i for all participants at all of the flow set points in a EURAMET low pressure gas flow comparison [7]. Figure 16 shows the results for the 4.5 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. The values of $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ are less than 0.96 for all participants and there are negligible differences between the $2u_{x_i}$, $2u_{d_i}$, and $2u_{\mathrm{lab}\ i}$ error bars. As we would expect for a comparison where $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$ are less than 1, all criteria deliver the same pass / fail results and there are no inconclusive results.

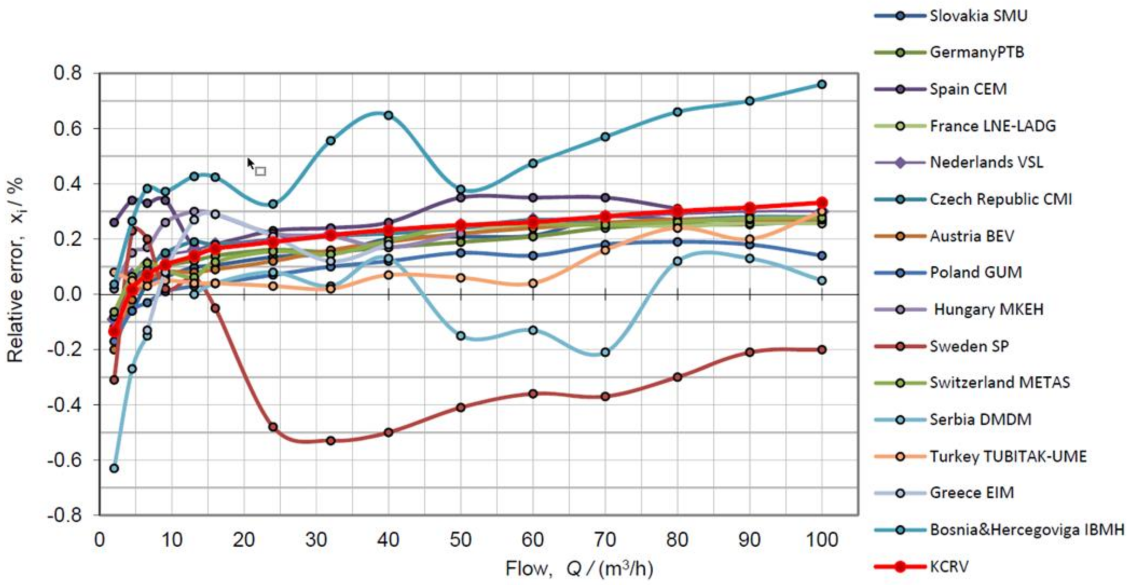


Figure 15. The CRV and reported results for a low-pressure gas flow comparison [7].

A) $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
B) $u_{\text{TS}}/u_{\text{lab}\ i} \leq 2$, $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
C) $L_{P,i} \leq 0.6$, etc.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
D) $P_i \geq 0.22$, etc.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓

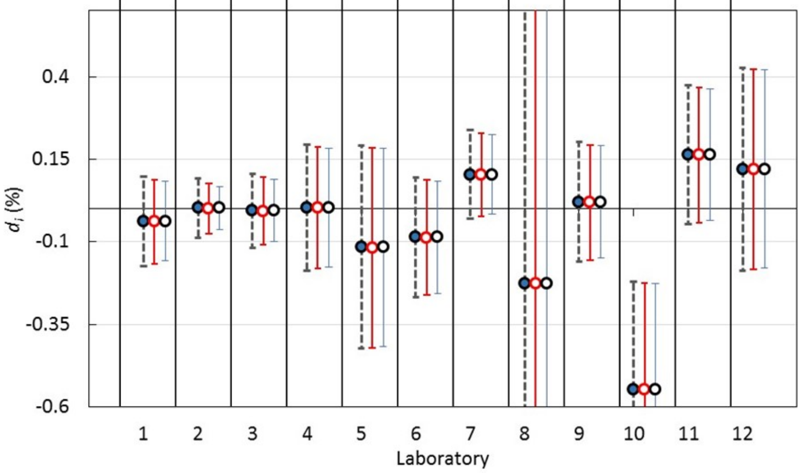


Figure 16. The PDFs, error bars, and criteria results for the 4.5 m^3/h set point for reference [7].

The next two data sets were selected because they have large $u_{\text{TS}}/u_{\text{lab}\ i}$ values. Figure 17 presents data from the 3.8 L/min set point of the hydrocarbon liquid flow proficiency test described in reference [13] (also shown in Figure 5). For this data set, $u_{\text{TS}}/u_{\text{lab}\ i}$ is greater than 2 (ranging from 2.2 to 5.7) and therefore, Criteria B reports all participant's results as inconclusive. Criteria C indicates two labs as inconclusive that the $\lvert En_i \rvert \leq 1$ criterion would call equivalent. Criteria D indicates four labs' results as inconclusive that the $\lvert En_i \rvert \leq 1$ criterion would call equivalent.

A) $\lvert En_i \rvert \leq 1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
B) $u_{\text{TS}}/u_{\text{lab}\ i} \leq 2$, $\lvert En_i \rvert \leq 1$	?	?	?	?	?	?	?	?	?	?	?	?
C) $L_{P,i} \leq 0.6$, etc.	✓	✓	✓	?	?	✓	✓	✓	✓	✓	✓	✓
D) $P_i \geq 0.22$, etc.	✓	✓	✓	?	?	?	✓	✓	✓	✓	?	✓

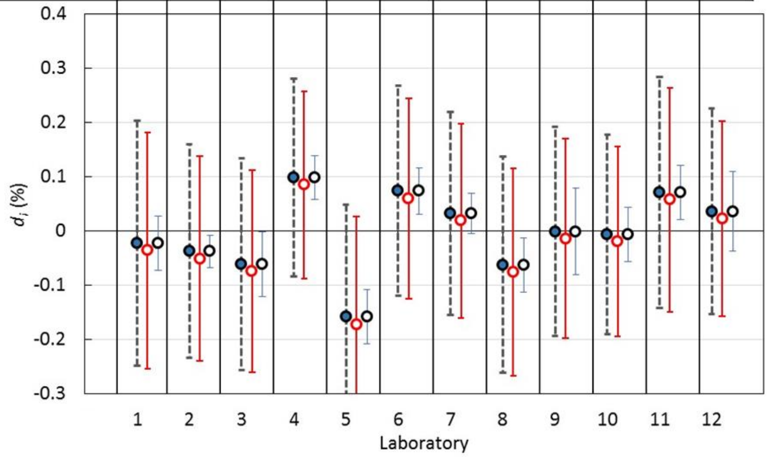


Figure 17. The PDFs, error bars, and criteria results for the 3.8 L/min set point for reference [13].

Figure 18 shows criteria results for the 40 m^3/h set point of a liquid flow comparison that had $u_{\text{TS}}/u_{\text{lab }i}$ greater than 2 (ranging from 1.8 to 4.7). Laboratory 7 is a case where the if $|d_i/(2u_{\text{lab }i})| \leq 1$ portion of Criteria D is not met, but the $P_i \geq P_{\text{th}}$ portion is met and therefore the participant passes.

A) $\lvert En_i \rvert \leq 1$	✓	✓	✓	✗	✓	✓	✓
B) $u_{\text{TS}}/u_{\text{lab }i} \leq 2$, $\lvert En_i \rvert \leq 1$	?	?	?	✗	✓	✓	?
C) $L_{P,i} \leq 0.6$, etc.	✓	?	✓	✗	✓	✓	✓
D) $P_i \geq 0.22$, etc.	✓	?	?	✗	✓	?	✓

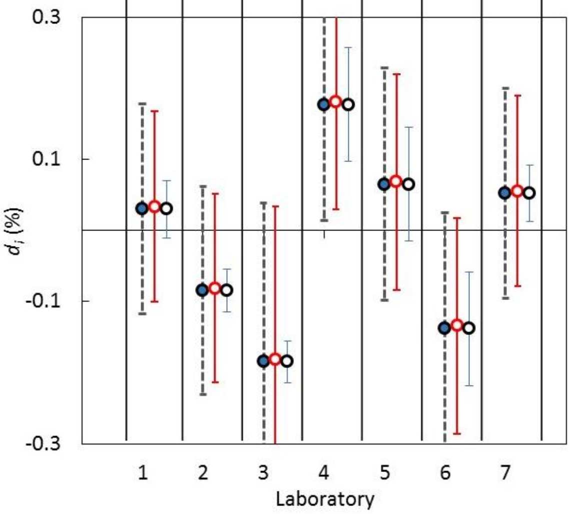


Figure 18. The PDFs, error bars, and criteria results for the 40 m^3/h set point for reference [15].

13. 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 shows that for $d_{\text{i}}/u_{\text{lab }i}$ and $u_{\text{TS}}/u_{\text{lab }i}$ > 2 (the northeast quadrant of Figure 6), the commonly used $|En_i| \leq 1$ criterion sometimes indicates equivalency when our visual assessment indicates the result should be 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. We demonstrated that calculating the maximal loss of power $L_{P,i}$ enables quantification of the deteriorating effect of a transfer standard's instability on the relevance of calculated degrees of equivalence. Criteria C uses the loss of explanatory power of the test $L_{P,i}$ to assess whether results should be considered inconclusive. We proposed to restrict $L_{P,i}$ to be no more than 0.6 for a laboratory to pass a comparison.

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 or lab uncertainty. Criteria D uses $|d_i/(2u_{\text{lab }i})|$, $|En_i|$, and P_i to mimic our visual assessment of comparison results. Criteria D passes labs that agree with the CRV within

$2u_{\mathrm{lab}\ i}$ even though that agreement may be due to random transfer standard errors. We applied the four criteria to the bi-lateral example and to real comparison data sets to better understand how they behave.

The threshold values for $L_{P,i}$ and P_i used in Criteria C and D are arbitrary and different values can be justified. For example, when we are working with 95 % confidence intervals: should there be a "warning level" to account for the other 5 % probability interval? Furthermore, it is important to find values that are appropriate for the particular measurand: a typical transfer standard for mass is more stable than one for liquid flow by orders of magnitude.

Criteria A gives inappropriate passing results for large $d_{\mathrm{i}}/u_{\mathrm{lab}\ i}$ and $u_{\mathrm{TS}}/u_{\mathrm{lab}\ i}$. Criteria B gives inconclusive results for some low $d_{\mathrm{i}}/u_{\mathrm{lab}\ i}$ values that we would consider passing. Criteria C and D successfully discerned between passing, failing, and inconclusive comparison results for the cases we have examined herein and future work should examine those criteria applied to more test cases.

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, and Yoshiya Terao.

[1] *Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008.

[2] *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.

[3] The BIPM Key Comparison Database, http://kcdb.bipm.org.

[4] Hartig, F., Bosse, H., and Krystek, M., *Recommendations for Unified Rules for Key Comparison Evaluation*, Key Engineering Materials, 613, pp. 26 to 33, 2014.

[5] *Calibration and Measurement Capabilities in the Context of the CIPM MRA*, CIPM MRA-D-04, Version 2, September 2010.

[6] *ILAC Policy for Uncertainty in Calibration*, ILAC-P14:12/2010.

[7] 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.

[8] Cox M. G., *Evaluation of key comparison data*, Metrologia, **39**, 589 to 595, 2002.

[9] 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.

[10] Rukhin, A. L., *Weighted Means Statistics in Interlaboratory Studies*, Metrologia, **46**, 323 to 331, 2009.

[11] 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**.**

[12] *Measurement Comparisons in the CIPM MRA*, CIPM MRA-D-05, version 1.5, March 2014.

[13] 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.

[14] 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, in press.

[15] 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.