FLOMEKO 2022

Evaluating Inter-laboratory Comparison Data

E. Frahm[1], J. Wright[2]

[1] *Physikalisch-Technische Bundesanstalt, enrico.frahm@ptb.de, Braunschweig, Germany*

[2] *National Institute of Standards and Technology, Gaithersburg, MD, USA*

E-mail (corresponding author): enrico.frahm@ptb.de

Abstract

The primary purpose of inter-laboratory comparisons is to demonstrate that the uncertainty specifications of the calibration measurement capabilities of the participating laboratories are correct. The most common criterion for assessing a participating laboratory's results is whether the normalized error $|En_i|$ is ≤ 1. Most comparison reports we reviewed properly include uncertainty components related to the transfer standard (u_{TS}) and the repeatability of the calibrations (u_{repeat_i}) in the uncertainty of the value reported by a participant. Unfortunately, high values for either u_{TS} and u_{repeat_i} decrease $|En_i|$, making it easier to achieve passing results in a comparison that uses a poor transfer standard or for a participant that delivers unstable measurements. A review of past comparison reports shows that this problem occurs for many measurands, including flow, temperature, and pressure.

Improved comparison criteria were proposed by [1] to counteract the flaws of the $|En_i| \leq 1$ criterion by introducing the possibility of inconclusive results and a probability-based approach. In this paper, we define comparison uncertainty u_{comp} as the root-sum-of-squares of u_{TS} and u_{repeat_i} and find it a better tool for assessing the power of the comparison than u_{TS} alone. We applied the comparison evaluation criteria to recent comparison results to illustrate their benefits over the $|En_i| \leq 1$ criterion. In general, the newer criteria confirm prior determinations, but in some cases passing results for the $|En_i| \leq 1$ criterion would be found inconclusive.

1. Introduction

The Working Groups of the Committee International des Poids et Mesure (CIPM) and National Metrology Institutes now have more than 20 years of experience of formal inter-laboratory comparisons to verify calibration measurement capabilities (CMCs). They have developed best practices related to evaluation of the performance and sensitivities of transfer standards, the necessary elements of a complete comparison report [2], and tools for processing the comparison data. The same methods are applicable to proficiency testing for ISO 17025 laboratory accreditation.

Despite the advances in methodology and the large effort involved in completing a comparison, the primary purpose of comparisons, *i.e.*, to determine whether the participants are meeting their uncertainty claims, remains sometimes vexing, subjective, and unsatisfactory. Unfortunately, the most commonly used metric, the normalized error $|En_i| \leq 1$ criterion has flaws, particularly if the comparison uncertainty u_{comp} is large relative to the reference standards being compared.

The CIPM Working Groups have discussed the topic since their inception, but the question remains: how should we apply the results of a comparison when assessing a proposed Calibration Measurement Capability?

Metrologists and statisticians responded to these questions, for instance [3]. We have learned how critical low uncertainty transfer standards are to a successful and conclusive comparison. More recently, [1] emphasized that in addition to showing that CMCs are valid (passing) or invalid (failing), comparison results can be inconclusive due to uncertainty introduced by the transfer standard and the comparison process. They also introduced several comparison criteria (including a probability-based criterion) that are easily applied in an Excel* spreadsheet. Malengo *et al.* [4] reviewed the history of this topic and proposed a probability-based criterion based on the previously developed statistics for assessing the conformity of an item to a specification.

In this paper, we review the currently available comparison criteria, show their value via some illustrative examples from recent comparisons, suggest an improvement in how the repeatability of comparison

* 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.

data is included in the analysis, and discuss application of these criteria to measurands other than flow, *e.g.*, mass, temperature, and pressure.

2. Review of evaluation approaches

2.1. Standard evaluation procedure: Criterion A

The commonly practiced evaluation procedure of key comparison data is based on Cox [3], also sometimes called Criterion A. The procedure in reference [3] calculates an uncertainty-weighted comparison reference value x_{CRV} and the degree of equivalence between participant i's reported value of the measurand x_i and the CRV: $d_i = x_i - x_{\mathrm{CRV}}$. The normalized error En for comparison participant i is:

$$En_i = d_i / 2u_{d_i}\,, \qquad (1)$$

where $2u_{d_i}$ is the 95 % confidence level uncertainty of the degree of equivalence.

By Criterion A, an $|En_i|$ value of $\leqslant 1$ indicates that the participant's result agrees with the CRV within the 95 % confidence level expectation. If the result exceeds the critical value of 1, the laboratory uncertainties to be verified are not confirmed by the comparison. The procedure includes a consistency check to exclude discrepant results by applying the chi-squared test [5]. Calculating En_i for each participating laboratory requires two input parameters: the reported value of the measurand x_i of laboratory i and the standard uncertainty of the reported value, u_{x_i}. Cox [3] did not elaborate on the components of u_{x_i} and some readers assumed that u_{x_i} included only the uncertainty of the participant`s reference standard, what we call u_{base_i} herein.

Many comparison pilots recognized that there are often significant contributions to the uncertainty of the participant's reported value other than u_{base_i}. When calculating u_{x_i}, they included additional uncertainties introduced by the transfer standard u_{TS}. They sometimes also included the repeatability of the reported value at each calibration point, expressed by s^2/n where s is the sample standard deviation and n is the number of repeated measurements. Here, we call the root-sum-of-squares of these values the comparison uncertainty u_{comp_i}:

$$u_{x_i} = \sqrt{u_{\mathrm{base}_i}^2 + u_{\mathrm{comp}_i}^2} = \sqrt{u_{\mathrm{base}_i}^2 + u_{\mathrm{TS}}^2 + s^2/n}\,. \qquad (2)$$

The transfer standard uncertainty u_{TS} includes all components introduced by the transfer standard and its associated instrumentation which could affect the measurement result during a comparison when the TS is used in a participant's lab. Often, the largest component of u_{TS} is drift or long-term calibration stability. Flow transfer standards are subject to calibration changes due to different fluid temperatures u_T, process pressures u_p and other property sensitivities u_{prop}:

$$u_{\mathrm{TS}} = \sqrt{u_{\mathrm{drift}}^2 + u_T^2 + u_p^2 + u_{\mathrm{prop}}^2 + \cdots}\,. \qquad (3)$$

In CCT-K2, a key comparison for standard platinum resistance thermometer calibrations [6] considered uncertainty introduced by thermal gradients, self-heating, drift correction, and other components related to the transfer standard that were not already included in the participant's calibration uncertainty analysis. Ideally, the comparison uncertainty would be negligible relative to the uncertainty of the participant's calibration standard. But for many measurands, limitations in the transfer standard preclude small values of $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ For some participants in CCT-K2, $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i} = 2.05$.

In CCM.P-K4.2012 [7], the drift and repeatability of the gauges used as the transfer standards in a pressure comparison were included in u_{x_i}. At the lowest pressure set point of 1 Pa, $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i} \cong 9$ for one of the participants.

The CIPM Working Group for Fluid Flow (WGFF) observed that some flow comparison uncertainties were large relative to the participants' CMCs and that large comparison uncertainty leads to smaller values of En. This raised the question of whether $|En_i| \leq 1$ was a sufficient criterion for assessing comparison results.

2.2. Consideration of comparison uncertainty and u_{comp_i}/u_{base_i}: Criterion B

The basic approach of [3] was adjusted by recommendations made by the WGFF in the field of flow and volume calibrations: [8] and [9]. In Criterion B, the ratio of the comparison uncertainty to the uncertainty of the participant's flow reference $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ is used as an additional evaluation criterion beside the En value. It is an indicator of whether the transfer standard is of sufficient quality to assess a participant's calibration capability. The WGFF proposed that $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i} \leq 2$ for conclusive comparison results and to avoid a participant passing solely because the transfer standard uncertainty and repeatability are large [1, 8, 10,]. Note that earlier publications used the ratio $u_{\mathrm{TS}}/u_{\mathrm{base}_i}$ not $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$but for reasons explained in section 3.2, we recommend $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ now.

In summary, Criterion B consists of two evaluation steps: 1) results with $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i} > 2$ are considered inconclusive and 2) for conclusive cases, the $|En_i| \leq 1$ criterion is applied to determine passing or failing results.

2.3. Consideration of probability density functions: Criterion D

Reference [1] introduced a probability-based criterion for the evaluating whether a comparison result is conclusive or not. This Bayesian approach assesses a probability in

view of the comparison data. The claims of the individual laboratories are evaluated by the degree to which two Gaussian probability density functions (PDFs) overlap, one representing the comparison reference value N(x_{CRV}, $u_{x_{\mathrm{CRV}}}$) and the other representing the participant's reported value for the measurand N(x_i, u_{base_i}). The degree of overlap is assessed by a probability content P_i of the CRV PDF bound by the participant's 2.5th and 97.5th percentile confidence limits for the uncertainty of the participant's flow reference. The Excel equation to calculate the probability P_i is given in [1] along with the recommended minimum or "threshold" value of P_{th} = 0.35. The threshold value determines the minimum required overlapping area between the PDFs of x_{CRV} and x_i [1] for a conclusive result. Section 4 shows examples of the PDFs and the area that probability P_i represents.

The probability-based approach behaves in a manner similar to our subjective evaluation of comparison results. For instance, if the TS or repeatability components in a comparison are large, the PDF for the CRV broadens and reduces P_i for participants that claim low uncertainty for their flow reference, unless their reported value coincides well with the CRV.

In summary, Criterion D consists of two evaluation steps: 1) results with $P_i > 0.35$ are considered inconclusive and 2) for conclusive cases, the $|En_i| \leq 1$ criterion determines passing or failing results.

Criteria B and D both apply a quality check of the comparison uncertainty to see if a result is conclusive before applying the $|En_i| \leq 1$ criterion.

3. Case studies for applying Criteria A and B

3.1. Importance of transfer standard uncertainty

The commonly used evaluation criterion $|En_i| \leq 1$ by [3] leads either to passing or failing comparison results. But [1] demonstrated that this approach is insufficient if the calibrations are influenced by the presence of transfer standard uncertainty u_{TS}, which can be quantified by Equation (3). The value of u_{TS} is a keystone for the interpretation of comparison results and for the definition of conclusive measurements. As an example, large values of u_{TS} would result in large values of $u_{\mathrm{x_i}}$ [Equation (2)]. This in turn would lead to smaller $|En_i|$ value and inaccurate CMC assessments if the basic approach of [3] is used alone. In such a case, a poor transfer standard with high uncertainty characteristics will produce low $|En_i|$ values. Although a laboratory would have passed the comparison, the results should actually be considered as inconclusive. To avoid such misinterpreted results and assuming the repeatability of the calibration is negligibly small, a maximum ratio of 2 for $u_{\mathrm{TS}}/u_{\mathrm{base}_i}$ was proposed by [8].

Two examples from recently completed comparisons are presented below to illustrate the problem.

During the COOMET.M.FF-S2 comparison [11], almost all participants successfully passed the $|En_i| \leq 1$ criterion (Figure 1a). In particular, laboratories 1 and 4 would have confirmed their CMC entries of $\leqslant$ 0.03 % very clearly. But at the same time, both laboratories also clearly exceeded the critical value of 2 for the ratio of $u_{\mathrm{TS}}/u_{\mathrm{base}_i}$ (Figure 1b). This effect was mainly caused by the long-term instability of the turbine meter transfer standard. Between the years 2009 and 2012, the TS showed a calibration drift of up to 0.23 % (Figure 1c). By applying Criterion B, the results for both laboratories were identified as inconclusive because the transfer meter was not suitable for a confirmation of the claimed CMC values.

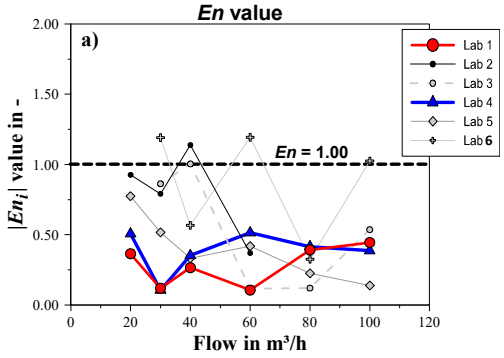


Figure 1a: Example of the importance of transfer meter uncertainty: acceptable values of $|\boldsymbol{En_i}| \leq \mathbf{1}$ for laboratories 1 and 4 *vs*. large values of $\boldsymbol{u_{\mathrm{TS}}/u_{\mathrm{base}_i}}$ in Figure 1b. Comparison results of COOMET.M.FF-S2 [11] for a turbine meter.

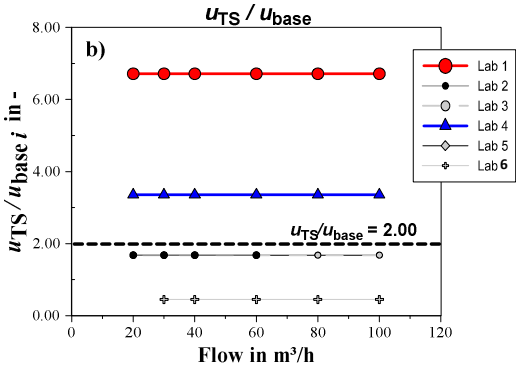


Figure 1b: Large values of $\boldsymbol{u_{\mathrm{TS}}/u_{\mathrm{base}_i}}$, caused by strong meter drift in Figure 1c, as an example for the importance of transfer meter uncertainty. Note that Labs 2, 4, and 5 have the same value of $\boldsymbol{u_{\mathrm{TS}}/u_{\mathrm{base}_i}}$ = 1.68 and their plots overlap.

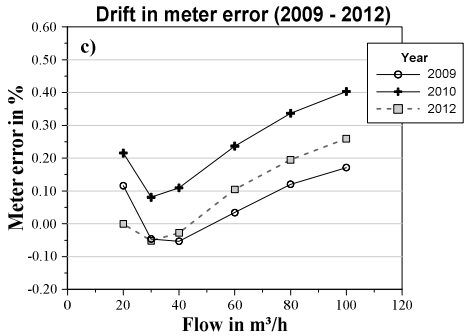


Figure 1c: Strong meter drift cause the large values of $\boldsymbol{u_{\mathrm{TS}}/u_{\mathrm{base}_i}}$ shown in Figure 1b.

The second example presents the results of a turbine meter which was used as the TS during key comparison CCM.FF-K1.2015 [12]. In this case, the meter sensitivity to disturbed inflow conditions (swirl) was identified as the main cause for inconclusive results (Figure 2c). All laboratories passed the $|En_i| \leq 1$ criterion (Figure 2a) but failed very clearly the $u_{\mathrm{TS}}/u_{\mathrm{base}_i}$ criterion by ratio values of up to 12 (Figure 2b). This example illustrates the importance of the additional discussion on conclusive and inconclusive results. Without considering the high sensitivity of the turbine meter to disturbed inflow conditions (Figure 2c), all laboratories would have successfully passed the comparison. But, following Figure 2b, the results of all participating labs had to be interpreted as inconclusive. Fortunately, a second, lower uncertainty transfer standard was also used in the comparison and provided conclusive results. Note that the turbine meter revealed valuable information about the velocity profile in the test sections of the participants.

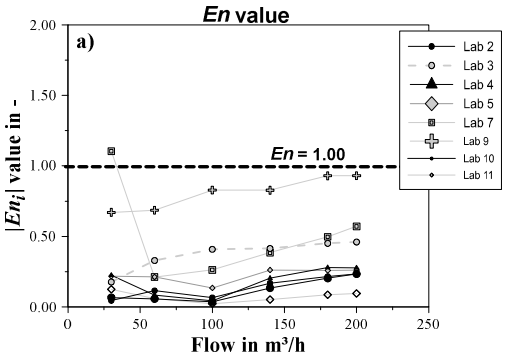


Figure 2a: Effect of large meter sensitivity on transfer standard uncertainty (Figure 2b) in contrast to acceptable values of $|\boldsymbol{En_i}| \leq \mathbf{1}$ criterion. Comparison results of CCM.FF-K1.2015 [12] for a turbine meter.

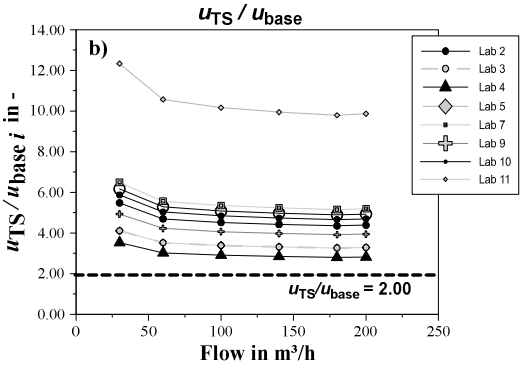


Figure 2b: Large transfer meter uncertainties $\boldsymbol{u_{\mathrm{TS}}}$ in contrast to acceptable values of $|\boldsymbol{En_i}| \leq \mathbf{1}$ criterion in Figure 2a.

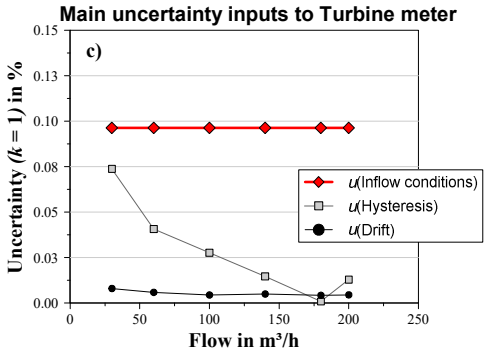


Figure 2c: Main inputs to the large amounts of transfer meter uncertainties in Figure 2b.

3.2. Importance of repeatability

Following Equation (2), the repeatability of calibrations is an essential input parameter to u_{x_i}. Like the previous discussion on the uncertainty of the transfer standard, high values of repeatability would lead to large values of u_{x_i} and finally, to reduced $|En_i|$ values.

The guideline [8] recommended that a laboratory's CMC value include an uncertainty component for the repeatability of the best existing device (BED). It can be assumed, that a transfer standard used in a key or supplementary comparison, would show comparable repeatability characteristics as a BED, but this is not always true. In consequence, in case of large values for u_{repeat_i}, this must be addressed with respect to the possibility of an underestimated uncertainty u_{base_i} of the participating laboratory. In the following two examples, we will present the importance of repeatability and which laboratories would be mainly affected by this discussion.

The first example presents preliminary results of a supplementary comparison (SIM.M.FF-S9-2016) [13] (Figure 3). Here, the evaluation is based on applying Equations (2) and (3) for a Coriolis flow meter transfer standard. Similar to previous discussions, the comparison would be successful if only the $|En_i| \leqslant 1$ criterion was used for data evaluation. But, at the lowest flow set point, laboratory 2 reported large values of repeatability (Figure 3c), leading to an acceptable $|En_i|$ value (Figure 3a). The need for an additional evaluation criterion is given by the large value of u_{repeat_i} (Figure 3c) which is followed by inconclusive measurements for the presented example due to a ratio of $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ larger than 2 at the lowest flow set point (Figure 3b). Note that this result would be considered conclusive if the repeatability were not included, *i.e.*, if the ratio $u_{\mathrm{TS}}/u_{\mathrm{base}_i}$ were used instead of $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ in Criterion B.

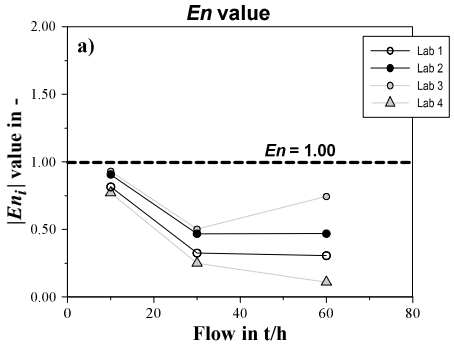


Figure 3a: Example of the importance of repeatability with all $|\boldsymbol{En_i}|$ values lower than 1 in contrast to unstable calibrations by Lab 2 (Figure 3b). Preliminary results of supplementary comparison SIM.M.FF-S9-2016 [13] of a Coriolis meter.

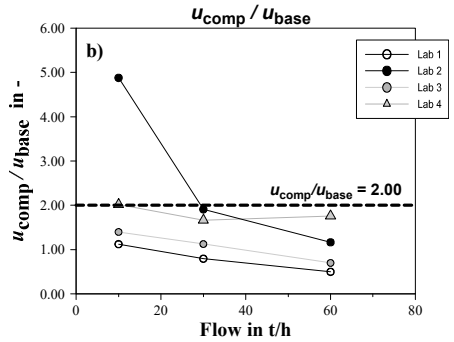


Figure 3b: Unstable calibrations by Lab 2 at lowest flow set point, expressed as $\boldsymbol{u_{comp_i}/u_{base_i}}$, caused by large values of $\boldsymbol{u_{repeat_i}}$ in Figure 3c, in contrast to acceptable values of $\boldsymbol{|En_i| \leq 1}$ in Figure 3a.

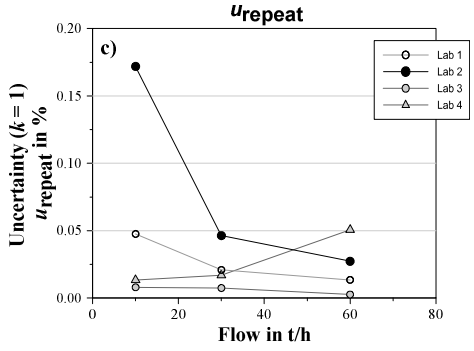


Figure 3c: Large values of $\boldsymbol{u_{repeat_i}}$ for unstable calibrations at lowest calibration point reported by Lab 2, which cause large values of $\boldsymbol{u_{comp_i}/u_{base_i}}$ in (Figure 3b).

The second example illustrates which laboratories would be mainly affected by producing large values for repeatability. The discussion is based on the COOMET.M.FF-S2 comparison [11]. Contrasting results from two laboratories were selected, one with a negligible amount of repeatability and a second laboratory where 50 % of the total uncertainty u_{x_i} was due to repeatability (Figure 4). For these original data, both laboratories successfully passed the $|En_i| \leq 1$ criterion as well as Criterion B (Table 1). In a second step, a fictitious set of data were produced by doubling the standard deviation of the repeated calibrations in Equation (2). The results in Figure 4 give no significant changes in the magnitude of u_{x_i} for laboratory 1 because of the low repeatability of its results. In contrast, the proportion of repeatability u_{repeat_i} for laboratory 2 increased to 80 % of u_{x_i}. At the same time, the $|En_i|$ value of laboratory 2 decreased to 0.79 but the ratio of $u_{\mathrm{comp}_i}/u_{\mathrm{base}_i}$ increased to values larger than 2 (Table 1).

In consequence, laboratories which do report very unstable results are more likely to show inconclusive comparison results than laboratories with low base uncertainties and low repeatability. Furthermore, without including the repeatability in Criterion B, large values of u_{repeat_i} reduce the $|En_i|$ value and produce passing results for laboratories with unstable calibration data.

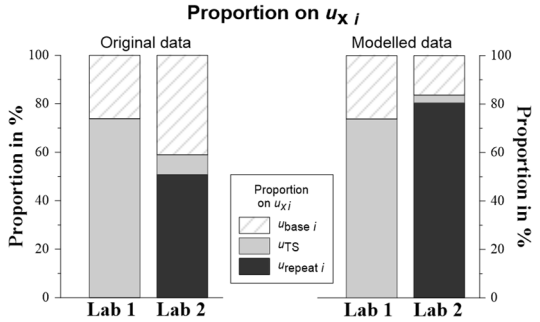


Figure 4: Proportion on uncertainty $\boldsymbol{u_{x_i}}$ for original and fictitious data (doubled the standard deviation used in Equation 2) for comparison results of COOMET.M.FF-S2 [11].

Table 1: Evaluation results of original and fictitious data (doubled standard deviation in Equation 2) for comparison COOMET.M.FF-S2 [11] in Figure 4.

Evaluation criterion	Original data		Modelled data	
	Lab 1	Lab 2	Lab 1	Lab 2
$\boldsymbol{En_i}$ value $\leqslant 1$	0.23	1.03	0.18	0.79
decision	passed	passed	passed	passed
$\boldsymbol{u_{comp_i}/u_{base_i}}$ $\leqslant 2$	1.68	1.20	1.68	2.27
decision	conclusive	conclusive	conclusive	in-conclusive

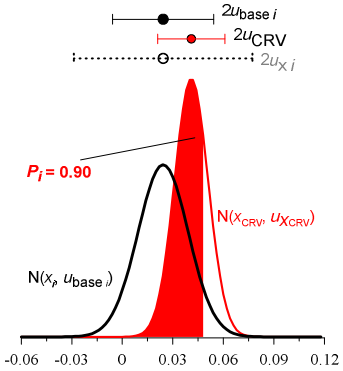
4. Probability-based evaluation - Criterion D

In this section, the probability-based Criterion D is applied to the data of COOMET.M.FF-S2, CCM.FF-K1.2015, and SIM.M.FF-S9-2016. No deviations in Criteria A ($|En_i| \leq 1$) and B were found between application of both criteria. In general, applying Criterion D to these comparisons gave similar results as Criteria A and B. In contrast, if Criterion D is additionally applied during the evaluation of comparison data, some assessments of CMC validity would be different. In the remainder of this section, four characteristic examples are discussed.

4.1. Example 1 - Passing and conclusive results

The first example demonstrates typical results of the re-evaluated comparisons. Laboratory i passed Criterion A and B (Figure 5). Also, the probability-based Criterion D

indicates conclusive results. There is no need for a reinterpretation of the final results: evaluations by Criteria A, B, and D give similar consequences for the comparison decision table and validation of CMC values.

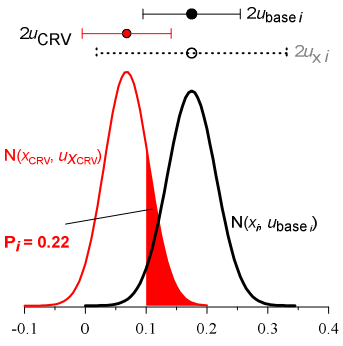


	Lab_i	x_{CRV}
x_i	0.024	0.041
d_i	-0.017	
$u_{base\,i}$	0.015	
u_{TS}	0.022	
$u_{comp\,i}$	0.022	
s^2/n	0.002	
$u_{x\,i}$	0.027	0.010
$u_{d\,i}$	0.025	
$d_i/u_{base\,i}$	-1.12	
En_i	0.34	Criterion A
$u_{comp\,i}/u_{base\,i}$	1.46	Criterion B
P_i	0.90	Criterion D

Figure 5: Representative results of a successful comparison evaluation where all criteria (A, B and D) indicate passing results that support the laboratory's CMCs. Results of re-evaluation for comparison data CCM.FF-K1.2015 [12] of a Coriolis meter. Besides the low $|En_i|$ value, the calibrations were conclusive due to significant overlapping areas ($P_i = 0.90$) of the PDFs N(x_{CRV}, $u_{x_{CRV}}$) and N(x_i, u_{base_i}).

4.2. Example 2 - Poor coincidence to CRV

The second example demonstrates the importance of introducing Criterion D to validate comparison results. In original evaluation, which was only based on Criteria A and B, participant i passed the comparison, *i.e.*, the $|En_i|$ value was 0.77 and the conclusiveness was verified by the ratio of $u_{comp_i}/u_{base_i} = 1.68$. But, if the probability-based Criterion D is additionally applied (Figure 6), the results of the comparison for laboratory i would be considered inconclusive. The overlapping area of the PDFs for x_{CRV} and x_i gives $P_i = 0.22$ (Figure 6), which is lower than the threshold value of 0.35 recommended by [1]. In this case, there is poor coincidence between the results of lab i and the CRV and the results are found to be inconclusive.



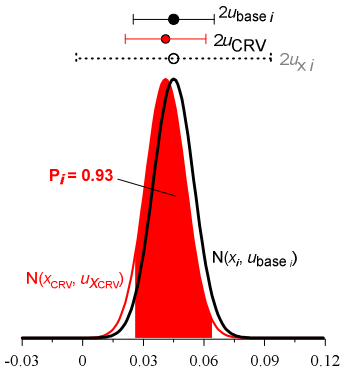
	Lab_i	x_{CRV}
x_i	0.175	0.068
d_i	0.107	
$u_{base\,i}$	0.04	
u_{TS}	0.067	
$u_{comp\,i}$	0.067	
s^2/n	0.002	
$u_{x\,i}$	0.078	0.037
$u_{d\,i}$	0.069	
$d_i/u_{base\,i}$	2.68	
En_i	0.77	Criterion A
$u_{comp\,i}/u_{base\,i}$	1.68	Criterion B
P_i	0.22	Criterion D

Figure 6: Example of inconclusive comparison results due to poor coincidence between laboratory results and the CRV. Results of re-evaluation for comparison data COOMET.M.FF-S2 [11] of a turbine meter. The PDFs N(x_{CRV}, $u_{x_{CRV}}$) and N(x_i, u_{base_i}) are presented ($P_i = 0.22$).

4.3. Example 3 - large transfer standard uncertainty

This example also demonstrates the importance of introducing Criterion D, but in the sense of supporting a participating laboratory with results that were originally considered inconclusive. In the original evaluation based on Criterion A, participant i easily passed the comparison with an $|En_i|$ value of 0.08. However, the results of the laboratory i were evaluated as inconclusive by Criterion B due to the large ratio of $u_{comp_i}/u_{base_i} = 2.19$. In contrast, in Figure 7 and the large overlapping area of the PDFs for x_{CRV} and x_i ($P_i = 0.93$) express very clearly the misinterpretation of the original evaluation.

For this example, the additional use of Criterion D gives the chance for laboratory i to pass the comparison successfully although Criterion B indicated the results were inconclusive. In this context, typical laboratories which were affected are characterized by a low lab uncertainty u_{base_i} in combination with a large transfer standard uncertainty u_{TS} (Figure 7).

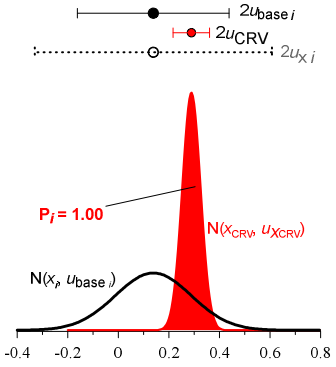


	Lab_i	x_{CRV}
x_i	0.045	0.041
d_i	0.004	
$u_{base\,i}$	0.010	
u_{TS}	0.022	
$u_{comp\,i}$	0.022	
s^2/n	0.002	
$u_{x\,i}$	0.024	0.01
$u_{d\,i}$	0.022	
$d_i/u_{base\,i}$	0.40	
En_i	0.08	Criterion A
$u_{comp\,i}/u_{base\,i}$	2.19	Criterion B
P_i	0.93	Criterion D

Figure 7: Impact of a large transfer standard uncertainty u_{TS} on the PDFs N(x_{CRV}, $u_{x_{CRV}}$) and N(x_i, u_{base_i}) if the uncertainty of laboratory i is comparatively small. Results of re-evaluation for comparison data CCM.FF-K1.2015 [12] for a Coriolis meter transfer standard.

4.4. Example 4 – large repeatability

This example demonstrates how Criterion D behaves when the repeatability is large. Large repeatability increases the uncertainty of the participants reported value, u_{x_i}, broadening the PDF for the CRV and reducing P_i unless the participant's reported value is coincident with the CRV. In the case shown in Figure 8, the PDF for the CRV N(x_{CRV}, $u_{x_{CRV}}$) falls entirely within the 95 % confidence interval of N(x_i, u_{base_i}) and therefore $P_i = 1.0$. Because $P_i \geq 0.35$ and $|En_i| \leq 1$, Criterion D indicates a conclusive and passing result for the participant's CMC claims.



	Lab_i	x_{CRV}
x_i	0.138	0.289
d_i	-0.151	
$u_{base\,i}$	0.150	
u_{TS}	0.067	
$u_{comp\,i}$	0.180	
s^2/n	0.167	
$u_{x\,i}$	0.234	0.036
$u_{d\,i}$	0.232	
$d_i/u_{base\,i}$	-1.00	
En_i	0.33	Criterion A
$u_{comp\,i}/u_{base\,i}$	1.20	Criterion B
P_i	1.00	Criterion D

Figure 8: Example for the influence of large amounts on repeatability on the PDFs N($\boldsymbol{x_{CRV}}$, $\boldsymbol{u_{x_{CRV}}}$) and N($\boldsymbol{x_i}$, $\boldsymbol{u_{base_i}}$). Results of re-evaluation for comparison data COOMET.M.FF-S2 [11] for a turbine meter.

5. Summary and conclusions

We introduced the root-sum-of-squares of the u_{TS} and u_{repeat_i} and called it the comparison uncertainty u_{comp_i}. Using u_{comp_i} instead of u_{TS} in Criterion B means that participants with poor repeatability are more likely to obtain an inconclusive result.

The sample of flow, mass, pressure, and temperature comparison reports we reviewed shows that the importance of quantifying the uncertainty introduced by the transfer standard is recognized across these measurands. The most often discussed component is drift (long term calibration stability) and it is assessed by multiple calibrations of the TS by the pilot lab before, during, and after the comparison. Our review shows that for some labs at some setpoints, it is not unusual to find that the uncertainty of the TS is significant enough to render some results inconclusive. We conclude that all measurands should begin applying more sophisticated criteria than $|En_i| \leq 1$.

The test cases presented in section 4 illustrate that probability-based Criterion D discerns cases that we consider are incorrectly interpreted as passing or failing by $|En_i| \leq 1$ alone or Criterion B.

Note that the criteria all have variable limits or threshold values related to the risk or confidence level desired. Even for the relatively simple $|En_i| \leq 1$ there is an ongoing discussion about whether or not to include a "warning level" when $1 < |En_i| \leq 1.2$. The values $u_{comp_i}/u_{base_i} \leq 2$ and $P_i > 0.35$ are suggestions that could be modified but seem to work well for the cases examined.

We encourage comparison pilot labs and organizations processing proficiency test data to utilize Criterion B or better still, Criterion D because they are both clear improvements over the $|En_i| \leq 1$ criterion and are easy to apply. A spreadsheet template that processes comparison data and implements the criteria is available from the authors or the WGFF upon request. We note that the guiding CIPM document [15] is flexible and allows for acceptance of calibration measurement capabilities based on evidence other than comparison results. Reviewers rely on their judgement when assessing CMCs and the comparison criteria described herein are tools that give pilot labs and assessors a more accurate view of comparison results by adding the inconclusive category.

6. Nomenclature

d_i	Degree of equivalence = $x_i - x_{CRV}$
En_i	Standardized degree of equivalence between a lab i and the key comparison reference value, = $d_i/2u_{d_i}$
i	Participating laboratory index
n	Number of measurements at one flow 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
s	standard deviation of a set of measurements, sample standard deviation
u_{base_i}	Type B standard uncertainty of the participating laboratory's reference standard
u_{comp_i}	standard comparison uncertainty, including transfer meter uncertainty and repeatability
u_{d_i}	Standard uncertainty of the degree of equivalence
u_{drift}	Standard uncertainty due to pressure sensitivity of the transfer standard
u_p	Standard uncertainty due to pressure sensitivity of the transfer standard
u_{prop}	Standard uncertainty due to property sensitivities of the transfer standard
u_{repeat_i}	Repeatability of measurements made by participant i, s^2/n, where s is the sample standard deviation and n is the number of measurements.
u_T	Standard uncertainty due to temperature sensitivity of the transfer standard
u_{TS}	Standard uncertainty of the transfer standard
$u_{x\,CRV}$	Standard uncertainty of the comparison reference value
u_{x_i}	Standard uncertainty of the reported value from the participating laboratory
x_{CRV}	Comparison reference value
x_i	Reported value of the measurand by the participating laboratory i

7. References

[1] Wright, J., Toman, B., Mickan, B., Wübbeler, G., Bodnar, O., Mickan, B., and Elster, C. (2016): Transfer standard uncertainty can cause inconclusive inter-laboratory comparisons. In: Metrologia 53, 1243 – 1258.

[2] BIPM Consultative Committee for Mass (2002): https://www.bipm.org/en/committees/cc/ccm.

[3] Cox M. G. (2002): Evaluation of key comparison data, Metrologia 39 589 – 595.
[4] Malengo, A., Bich, W. (2022): Conformance probability in the assessment of Calibration and Measurement Capabilities. Measurement, 192, 110865
[5] Cox M. G. (2007): the evaluation of key comparison data: determining the largest consistent subset, Metrologia, 44, 187 – 200
[6] Steele, A. G. *et al*. (2002): CCT-K2: key comparison of capsule-type standard platinum resistance thermometers from 13.8 K to 273.16 K, Metrologia, 39, pp. 551.
[7] Ricker, J., et al., (2017): Final report on the Key Comparison CCM.P-K4.2012 in absolute pressure from 1 kPa to 10 kPa, Metrologia, 54, Tech. Suppl., 07002.
[8] WGFF (2013): WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty. 21.10.2013, http://www.bipm.org/utils/en/pdf/ccm - wgff - guidelines.pdf.
[9] Wright, J., Mickan, B., Benkova, M., Chun-Lin Ch. (2022): Laboratory Comparison Calculations Spreadsheet Template Following Cox 2002 Method, BIPM WGFF, unpublished working paper.
[10] Wübbeler, G., Bodnar, O., Mickan, B. and Elster, C. (2015): Explanatory power of degrees of equivalence in the presence of a random instability of the common measurand. Metrologia 52, 400–5.
[11] Frahm, E., *et al*., (2020): Final report on supplementary comparison of national standards for liquid flow COOMET.M.FF-S2 (COOMET Project 406/UA/07), Metrologia, 57, Tech. Suppl., 07023.
[12] Frahm, E., Furuichi, N., Arias, R., Yang, C.-T., Chun, S., Meng, T., Shinder, I., Büker, O., Mill, Chr., Akselli, B., Smits, E. (2022): Final report on Key Comparison CCM.FF-K1.2015. In final review.
[13] Frahm, E. *et al*. (2022): Final report on Supplementary Comparison SIM.M.FF-S9. Report in progress, Draft A.
[14] Chiang, Ch.-L. *et al.* (2022): Final report on CCM.FF-K6.2017. Report in progress, Draft B.
[15] Calibration and Measurement Capabilities in the context of the CIPM MRA - CIPM MRA-D-04, January 2011, page 13.