Laboratory Comparison Calculations Spreadsheet Template Following Cox 2002 Method

John Wright, Bodo Mickan, Miraslova Benkova, Chun-Lin Chiang

April 14, 2022

Nomenclature

d_i	Degree of equivalence $= x_i - x_{\mathrm{CRV}}$.
En_i	Standardized degree of equivalence between a lab i and the comparison reference value, $= d_i/2u_{d_i}$.
i, j	Participating lab index.
k	Coverage factor associated with a specified confidence level.
$\mathrm{N}(\mu, \sigma)$	Normal, Gaussian probability distribution with mean μ and standard deviation σ.
n	Number of measurements made at a set point, or number of participants, components, etc.
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), including uncertainty for the best existing device.
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 [1] excluding uncertainty for the best existing device.
$u_T, u_P, u_{\mathrm{prop}}$	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.
$u_{\mathrm{comp}\,i}$	$u_{\mathrm{comp}\,i} = \sqrt{{u_{\mathrm{TS}}}^2 + {u_{\mathrm{R},i}}^2}$ where $u_{\mathrm{R},i}$ is the repeatability measured in each participant's lab.
$u_{\mathrm{R}\,i}$	Repeatability, quantified by the standard deviation of the mean (also called standard error), *i.e.,* the sample standard deviation divided by the square root of the number of measurements n.
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.
U	Expanded, approximately 95 % confidence level uncertainty.

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

x_i	Reported value of the measurand by the participating laboratory i.
x_{CRV}	The comparison reference value.
χ_ν^2	Chi-squared statistic for $\nu = n - 1$ degrees of freedom. Used for identifying discrepant values via the Chi-squared consistency check.
ν	Number of degrees of freedom.

Introduction

The purpose of this document is to describe the processing of laboratory comparison data and assessing whether participants' results agree with each other within uncertainty expectations (passing, failing, or inconclusive). It provides an overview and some instructions on how to use a spreadsheet called "Comparison Calculation Temple Day-Month-Year.xlsx".

Participants' Data

The Pilot laboratory should provide a spreadsheet template for the electronic submission of data by each participating lab. The spreadsheet should include: 1) the n measurements of the measurand at each set point, 2) the base uncertainty of the participant's reference standard u_{base}. An alternative list is: 1) the *average* of n measurements of the measurand at each set point, 2) the *standard deviation of the mean*[2] of the measurand, and 3) the base uncertainty of the reference standard u_{base}.

A flow meter will be the transfer standard for most WGFF comparisons. A commonly used value of the measurand from the participating laboratory (x_i) is the percent difference between the flow indicated by the transfer standard and the participant's reference standard. But dimensional values work equally well and the measurand x_i might have units of L for a volume comparison or kg for a mass comparison.

See the WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty[3], for an explanation of u_{base} and note that it does not include the repeatability of the best existing device. Also note that the Pilot may need more data reported by the participants than listed above, it is a minimum necessary data set. Examples of other data that might be useful are fluid temperature, pressure, room conditions. These variables might be used to evaluate uncertainties in the transfer standard due to temperature sensitivity etc.

Uncertainty of the Reported Value

The methods published by Cox[4] are commonly used for calculations of the Comparison Reference Value (CRV), its uncertainty, and the uncertainty of the degree of equivalence. The uncertainty of the reported value (u_{xi}) is ***not*** simply the uncertainty of the participant's flow reference (u_{base}): it also includes uncertainties introduced by the transfer standard and repeatability of the reported value at each set point:

---

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

3 *WGFF Guidelines for CMC Uncertainty and Calibration Report Uncertainty*, October 21, 2013 http://www.bipm.org/en/committees/cc/ccm/working_groups.html.

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

$$u_{xi} = \sqrt{u_{\text{base}}^2 + u_{\text{TS}}^2 + \frac{s^2}{n}} \quad (1)$$

The transfer standard uncertainty is the root-sum-of-squares (RSS) of the calibration drift of the transfer standard (and its associated instrumentation) during the comparison, temperature sensitivities, pressure sensitivities, property sensitivities, and perhaps other components:

$$u_{\text{TS}} = \sqrt{u_{\text{drift}}^2 + u_{\text{T}}^2 + u_{\text{P}}^2 + u_{\text{prop}}^2 + \cdots} \quad (2)$$

The uncertainty due to calibration drift is normally quantified by performing repeated calibrations in the Pilot lab using the same reference standard before, during, and after the comparison. If a large enough number of calibrations in the Pilot lab are available ($n \geq 5$), the standard deviation s is a good measure of long-term calibration drift from unknown sources. For $n < 5$, a rectangular distribution can be applied to the range of the calibration changes observed in these Pilot lab calibrations $(x_{\text{max}} - x_{\text{min}})$:

$$u_{\text{drift}} = \frac{(x_{\text{max}} - x_{\text{min}})}{2\sqrt{3}} \quad (3)$$

A similar approach is used to assess other transfer standard sensitivities. For example, during preliminary testing in the Pilot lab, the transfer standard can be placed in a temperature-controlled chamber and exposed to a range of environmental and fluid temperatures. These experiments are used to make corrections for the participants' lab conditions or to establish a range of conditions that are acceptable within the participating labs necessary to hold uncertainties due to temperature to an acceptable level. In cases where the fluid properties or pressure can be reasonably expected to influence a transfer standard, uncertainty components should be included for those components. A checklist of possible transfer standard sensitivities is available on the WGFF restricted web page.

Processing Comparison Data

Once the transfer standard uncertainty is quantified, it is combined by RSS with each participant's base uncertainty and the standard deviation of the mean of the n measurements at each flow set point (see Equation 1). At this point we have the uncertainty of the reported values, u_{xi}, (different for each participant as indicated by the lab index i) and the procedures given by Cox[4] are reproduced below.

The comparison reference value, x_{CRV} is an average of the participant's results that gives greater weighting to participants with lower uncertainty:

$$x_{\text{CRV}} = \frac{\frac{x_1}{u_{x_1}^2} + \frac{x_2}{u_{x_2}^2} + \cdots + \frac{x_n}{u_{x_n}^2}}{\frac{1}{u_{x_1}^2} + \frac{1}{u_{x_2}^2} + \cdots + \frac{1}{u_{x_n}^2}} \quad , \quad (4)$$

where x_1, x_2, ... x_n are values of the measurand at one set point in different independent laboratories 1, 2, ...n and u_{x_1}, u_{x_2}, ... u_{x_n} are standard uncertainties of the reported values for participating laboratories 1, 2, … n (including the uncertainty introduced by the transfer standard and the standard deviation of the mean for n measurements made at each set point.

The participants' uncertainty analyses should be reasonable, neither too pessimistic nor aspirational or the CRV can be skewed towards an unrealistic value. The standard uncertainty of the reference value $u_{x\ \mathrm{CRV}}$ is given by:

$$\frac{1}{u_{x_{\mathrm{CRV}}}^2} = \frac{1}{u_{x_1}^2} + \frac{1}{u_{x_2}^2} + \cdots + \frac{1}{u_{x_n}^2} \,. \tag{5}$$

The expanded uncertainty of the reference value is:

$$U_{x\ \mathrm{CRV}} = 2\, u_{x\ \mathrm{CRV}} \,. \tag{6}$$

The Chi-squared consistency check finds discrepant results by comparing expected and observed values of the Chi-squared statistic at each comparison set point. The observed Chi-squared value χ_{obs}^2 is calculated from the comparison data:

$$\chi_{\mathrm{obs}}^2 = \frac{(x_1 - x_{\mathrm{CRV}})^2}{u_{x_1}^2} + \frac{(x_2 - x_{\mathrm{CRV}})^2}{u_{x_2}^2} + \cdots + \frac{(x_n - x_{\mathrm{CRV}})^2}{u_{x_n}^2} \,. \tag{7}$$

The expected Chi squared value depends on the confidence level and the degrees of freedom ν:

$$\nu = n - 1 \,, \tag{8}$$

where n is a number of participating laboratories.

If the observed Chi-squared value is smaller than the expected Chi-squared value χ_{υ}^2 for the 95 % probability level and the number of degrees of freedom, *i.e.,* if $\chi_{\mathrm{obs}}^2 < \chi_{\upsilon}^2$, then the data set passes the Chi-squared consistency check and the x_{CRV} and $U_{x\ \mathrm{CRV}}$ values are accepted.

If $\chi_{\mathrm{obs}}^2 > \chi_{\upsilon}^2$, then the Chi-squared consistency check fails and the laboratory with the highest value of $\frac{(x_i - x_{\mathrm{CRV}})^2}{u_{x_i}^2}$ is excluded from a new iteration of calculations of x_{CRV}, $U_{x\ \mathrm{CRV}}$, χ_{obs}^2, ν, and χ_{υ}^2. This process is repeated until the remaining laboratories pass the consistency check. If a large fraction of the participants fail the Chi-squared consistency check, other approaches for calculating the CRV (like the median) should be considered.

1.1 Degree of equivalence

After the CRV is determined, the degree of equivalence (the differences between each participating laboratory's result and the CRV) is calculated according to

$$d_i = x_i - x_{\mathrm{CRV}} \,. \tag{9}$$

Based on these differences, the standardized degree of equivalence En is calculated according to:

$$En_i = \frac{d_i}{2u_{d_i}}, \tag{10}$$

In general, $|En_i| \le 1$ indicates that a participating laboratory agrees with the CRV within uncertainty expectations.

The uncertainty of the degree of equivalence u_{d_i} is calculated by one of two formulas that depend on whether or not participant i's result x_i is correlated with the comparison reference value x_{CRV}.

Recall the equation for combined uncertainty with correlated components (Equation 13 of the GUM[5]):

$$u_c^2(y) = \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j} u(x_i, x_j).$$

For the degree of equivalence (Equation 9), this simplifies to:

$$u_{d_i}^2 = u_{x_i}^2 + u_{\mathrm{x\ CRV}}^2 - 2u_{x_i,\mathrm{x\ CRV}} = u_{x_i}^2 + u_{\mathrm{x\ CRV}}^2 - 2r_{x_i,\mathrm{x\ CRV}} u_{\mathrm{x\ CRV}}^2 \tag{11}$$

Where $u_{x_i,\mathrm{x\ CRV}}$ represents the covariance between the CRV and the participant's measurement and $-1 \le r_{x_i,\mathrm{x\ CRV}} \le 1$ is the correlation coefficient. If a participant's measurement was included in the calculation of the CRV, x_i and $\mathrm{x}_{\mathrm{CRV}}$ are correlated and $r_{x_i,\mathrm{x\ CRV}} = 1$. Hence for most comparison participants*,* Equation 11 simplifies to:

$$u_{d_i} = \sqrt{u_{x_i}^2 - u_{\mathrm{x\ CRV}}^2}\,. \tag{12}$$

Equation 12 applies to two types of participants: 1) the participants that contributed to the CRV calculation and 2) the participants that are traceable to another participant's reference standards and therefore are not an independent realization of the unit. Dependent participants' results are assumed to be correlated with the CRV even though the measurements they submitted were not used in the CRV calculation. An example of dependent traceability is: Lab 1 calibrates reference flow meters for Lab 2 and Lab 2 uses those flow meters in the comparison. If the two lab's measurements were fully correlated ($r_{1,2} = 1$, covariance $2u_{x_i,\mathrm{x\ CRV}} = 0$), including both labs' results in the CRV calculation would double the weight of Lab 1 in the CRV (inappropriate). Equation 12 appears as Equation 5 in Cox 2002 and applies to most comparison participants.

The uncertainty of the degree of equivalence for uncorrelated participants is:

$$u_{d_i} = \sqrt{u_{x_i}^2 + u_{\mathrm{x\ CRV}}^2}\,. \tag{13}$$

[5] *Guide to the Expression of Uncertainty in Measurement*, JCGM 100:2008, page 21.

Equation 13 applies when the participant's results were excluded from the CRV calculation because: 1) their results were determined to be outliers by a statistical test or 2) in an international key comparison, the participant is not a signatory of the Mutual Recognition Arrangement or a Designated Institute.

Note that Equations 12 and 13 apply to methods that use a weighted mean of the participants' results as the CRV. If the median of the participants' results is used as the CRV (Cox's Procedure B), a Monte Carlo simulation is used to determine u_{d_i}. The comparison spreadsheet template does not presently have an option that applies Procedure B.

The expanded uncertainties U_{d_i} are determined by using a coverage factor of $k = 2$ to obtain an approximately 95 % confidence level value:

$$U_{d_i} = 2\, u_{d_i}\ . \tag{13}$$

Three methods of assessing equivalence are applied in the spreadsheet template[6].

- Criterion A assesses participants' results by: Participant i passes if $|En_i| \leq 1$ and fails if $|En_i| > 1$.
- Criterion B requires that the ratio of the "comparison uncertainty" u_{comp} and participant's base uncertainty be below some threshold: Participant i passes if $u_{\mathrm{comp}}/u_{\mathrm{base}\,i} \leq R_{\mathrm{th}}$ and $|En_i| \leq 1$, fails if $|En_i| > 1$, and the comparison results are inconclusive for participant i otherwise. The spreadsheet template uses $R_{\mathrm{th}} = 2$. Here, comparison uncertainty is defined as $u_{\mathrm{comp}} = \sqrt{{u_{\mathrm{TS}}}^2 + {u_{\mathrm{R},i}}^2}$ where $u_{\mathrm{R},i}$ is the repeatability measured in each participant's lab.
- A probability-based criterion is also applied in the spreadsheet template, 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. The spreadsheet template uses $P_{\mathrm{th}} = 0.35$.

Note: According to the 14th CCM meeting (February, 2013) pair-wise degrees of equivalence should no longer be published in the COMPARISONDB. Information on pair-wise degrees of equivalence published in COMPARISON reports should be limited to the equations needed to calculate them, with the addition of any information on correlations that may be necessary to estimate them more accurately.

Notes from Excel Spreadsheet template:

Documentation worksheet:

Date	Author	Change to spreadsheet
5-Dec-11	J. Barbe	First version to come into J. Wright's possession from Jean Barbe of LNE.

[6] Wright, J., Toman, B., Mickan, B., Wubbeler, G., Bodnar, O., and Elster, C., *Transfer Standard Uncertainty Can Cause Inconclusive Inter-Laboratory Comparisons*, Metrologia 53, pp. 1243 to 1258, October, 2016.

21-Jan-16	J. Wright	Added degree of equivalence plots and tables showing results from Criteria A, B, and D that are explained in Wright, J., Toman, B., Mickan, B., Wubbeler, G., Bodnar, O., and Elster, C., *Transfer Standard Uncertainty Can Cause Inconclusive Inter-Laboratory Comparisons*, Metrologia 53, pp. 1243 to 1258, October, 2016.
11-Mar-20	J. Wright	User can remove discrepant results by entering negative values of Ubase in Table 2.
12-Jun-20	J. Wright	Added conditional formatting so that discrepant results turn red in column AB. The user should then change the participant's uncertainty for the discrepant measurement to a negative value in the appropriate cell in Table 2. This removes that measurement from the CRV calculation and uses the uncorrelated equation for calculating u(di). This spreadsheet was validated by comparison of results from an independently programmed spreadsheet made by Bodo Mickan and/or Rainer Kramer of PTB called CCM-K6-2017_RV-calculations.xlsx when applied to the K6-2017 gas flow comparison.
1-Sep-20	J. Wright	Added traceability dependence (Table 4) where +1 or -1 is entered depending upon whether the lab's flow traceability is independent (+1) or dependent (-1) of other participants. These values are used during calculation of the uncertainty of the degree of equivalence and also in uncertainty weighting (columns X and Y), Chi-squared calculations (AA and AB). Added columns for entry of suggested transfer standard uncertainty components (root-sum-squared in column G).
6-Oct-20	J. Wright	Fixed conditional formatting for column W, Lab J. Fixed criterion D formulas.
15-Apr-21	J. Wright	Improved graph at column CF.
14-Apr-22	J. Wright	Changed the numerator in Criterion B to use $U_{\mathrm{comp}\ i}/U_{\mathrm{base},i}$ instead of $U_{\mathrm{TS}}/U_{\mathrm{base},i}$ where U_{comp} is the RSS of the transfer standard uncertainty (U_{TS}) and the repeatability measured in each participating lab ($U_{\mathrm{R},i}$). This change was made based on suggestions from Frahm and Buker.

Text and Notes from Cells in Comparison calcs worksheet:

B1: This spreadsheet will calculate a comparison reference value (CRV), degrees of equivalence, uncertainties, etc. following the methods of Cox 2002.

B2: It will handle up to 15 labs and 18 set points: use [-] in unused cells. Cells shaded yellow can be used as inputs, the rest are outputs.

B3: The user can remove discrepant results by entering negative values of Ubase in Table 2. Chi-squared test is applied at cell U321:U339.

B4: Look for notes on individual cells and see the document called "Laboratory Comparison Calculations Spreadsheet Template Following Cox 2002 Method" for instructions on using this spreadsheet.

B6: This table allows the user to enter transfer standard uncertainty components and sensitivities related to the transfer standard, e.g. calibration stability, property effects, sensitivity to temperature, etc. Column G is the RSS of those components.

C7: UT (k=2, %), Transfer standard uncertainty due to temperature sensitivity and the temperature differences between participants' labs.

D7: UP (k=2, %), Transfer standard uncertainty due to pressure sensitivity.

E7: Uother (k=2, %), Transfer standard uncertainty due to other components, i.e. differences in fluid properties between participants' labs, leaks, etc

F7: UR (k=2, %), Long term calibration stability of the transfer standard usually quantified by repeated calibrations performed in the pilot lab.

G7: UTS (k=2, %), This is the root-sum-of-squares of transfer standard uncertainty components.

B28: Table 1. Reported relative error of the TS, x (%), Error of the transfer standard relative to the participating laboratory's reference standard, in percent.

B50: Table 2. Reported participating lab uncertainty, Ubase, i , k=2 (%), k = 2 uncertainty, in %, of the participating lab's reference standard without uncertainty due to the best existing device (BED) or transfer standard reproducibility.

Use a negative uncertainty value if 1) the participant's value is found discrepant by the CHi-squared test or 2) if this is a Key Comparison and the participant is not an MRA signatory or Designated Institute.B72: Table 3. Reproducibility, UR , k=2 (%), k = 2 reproducibility, in %, = (2 s) / (xbar sqrt (n)) 100 for the n repeated measurements of the transfer standard in the participant's lab.

B72: Table 3. Reproducibility, UR , k=2 (%), k = 2 reproducibility, in %, = (2 s) / (xbar sqrt (n)) 100 for the n repeated measurements of the transfer standard in the participant's lab.

B94: Table 4. Participating lab's traceability dependence: 1=independent of other participants, -1=dependent on another participant., Enter 1 for labs that have independent traceability. Enter -1 for labs with dependent traceability.

B116: Table 5. Uncertainty of the reported value Ux i = RSS [Ubase, UTS, UR], k=2 (%)., Optional Input!!, k = 2 uncertainty of the reported value, in %, should be the RSS of base unc., transfer standard unc., and reproducibility in participant's lab… but it is better to input the values in tables 0, 2, and 3 and use the RSS formula here so you can see how big the values are in Table 5.

B138 Table 6. Ucomp / Ubase, i, Ucomp is the RSS of UTS and UR measured in the participating lab.

B161: Table 7. Degree of equivalence, di (%).

B183: Table 8. Uncertainty of degree of equivalence, U(di) (k=2, %).

B205: Table 9. En values

U7: Comparison Reference Value (CRV) calculations

T28: Lab Count, The lab count is the number of participants at each set point that do not have discrepant results or dependent traceability.

V28: CRV, comparison reference value.

W28, X28: K=1 and k=2 uncertainty of the CRV.

AB51: Conditional formatting highlights cells in this column red if they are a discrepant result. Such results should be removed from the CRV calculation by entering a negative uncertainty value in the appropriate cell in Table 2.

Z52, AB52, AC52, Ad52, AE52: Don't drag this column's cells indiscriminately!

W52: Conditional formatting turns these cells red if the value has been removed by entering a negative uncertainty in Table 2 or -1 in Table4.

V325 to V342: If this column says OOPs, the chi squared consistency check did not pass for one or more labs.

AL5: Results by set point.

BD5: This section handles only 10 labs so far (not 15).