Guide to the Analysis of CCT Key Comparison Results

Prepared by the CCT Working Group for Key Comparisons

October 2026

1. Introduction

This document provides guidance on the analysis of results of CCT Key Comparisons (KCs). Since different approaches are followed by other CIPM Consultative Committees, this document is applicable only to CCT KCs.

The analysis of the results of a KC includes many steps and, in this document, these steps are discussed in the same chronological order in which the KC pilot and the participants have to take them (see Figure 1).

Given the terseness and ambiguity of the information of CIPM MRA documents on the key comparison reference value (KCRV) and the degrees of equivalence (DoEs), we deemed necessary a clarification of the definitions and interpretations of the KCRV and DoEs in CCT KCs as the first and preliminary step (**Section 2**).

Moreover, because some decisions about the methods of analysis need to be taken when designing the KC Technical Protocol, we discuss the general principles to be followed in the analysis that should be set in the Technical Protocol (**Section 3**).

Section 4 describes the inspection of the initial data set using simple graphical methods. Such inspection provides a qualitative assessment of the dispersion of the results and allows an identification of anomalous results.

Section 5 provides guidance on how to deal with anomalous results, amendment and withdrawal of results, and selection of the final data set to be used in the analysis.

Section 6 describes the statistical tests to be performed on the final data set to obtain useful information for the following analysis.

Section 7 describes how bilateral degrees of equivalence can be directly computed from the final data set, without needing to select a procedure for computing the KCRV (and without computing the KCRV), provided correlations and dark uncertainty (if detected) are properly taken into account.

Section 8 recalls the four commonly used statistical models in key comparisons and describes the random effects model, which is considered the most appropriate statistical model for CCT KCs.

Section 9 summarizes the most commonly used statistical procedures for computing the KCRV.

Finally, **Section 10** provides guidance on the selection of the statistical procedure to be used for computing the KCRV.

This document draws inspiration from two major sources: 1) the CCQM Guidance note "Estimation of a consensus KCRV and associated Degrees of Equivalence" [1], which is currently under revision by a CCQM Task Group, and 2) the NIST Decision Tree User's Manual (3rd Edition) and its background article, see Appendix I and [2].

This document is not intended to be prescriptive or mechanically applied to all CCT KCs. Rather, it provides a useful framework to allow the pilots and participants to make the appropriate decision on the selection of the method to be used in the determination of the KCRV, basing it on their expertise, measurement science and statistical science.

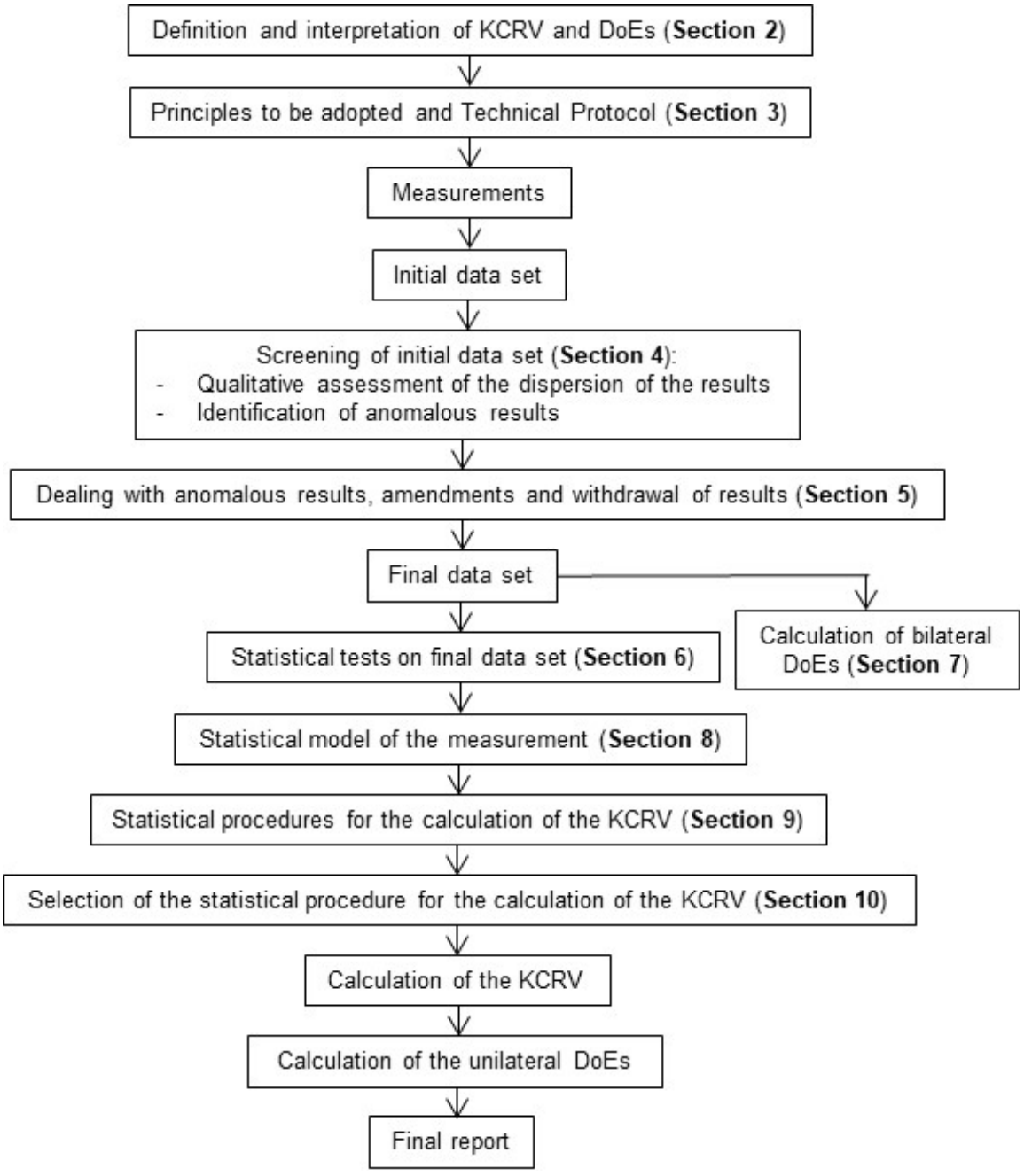


Figure 1: Flow chart of a CCT Key Comparison. The sections of this document mirror this flow chart.

2. Degrees of Equivalence and Key Comparison Reference Value

In the framework of the CIPM MRA [3], the primary objective of a Key Comparison (KC) is to obtain a quantitative measure of the degrees of equivalence (DoEs) of national measurement standards maintained by the national metrology institutes (NMIs). The MRA defines (though merely in words, without using any formula) two forms of DoEs, often referred to as unilateral DoEs and bilateral DoEs. The definition of DoEs involves the key comparison reference value (KCRV) and the meaning to be attributed to it. However, the information provided by the MRA documents (essentially, [3] and [4]) on DoEs and KCRV is terse, ambiguous and under-determined, so in this section we clarify how the CCT interprets the DoEs and the KCRV.

2.1 Degrees of Equivalence in CCT Key Comparisons

2.1.1 Unilateral Degrees of Equivalence

The unilateral DoE, D_i, is the degree to which the standard of participant i is consistent with the KCRV. Indicating with x_i the value measured by participant i in the KC and with x_{KCRV} the value of the KCRV, D_i is quantitatively expressed by the two following terms:

$$\begin{aligned} D_i &\equiv x_i - x_{\mathrm{KCRV}} \\ U(D_i) &\equiv U(x_i - x_{\mathrm{KCRV}}) \end{aligned}$$

where $U(D_i)$ is the expanded uncertainty of D_i at a 95 % level of confidence.

2.1.2 Bilateral Degrees of Equivalence

The bilateral (or pairwise) DoE, D_{ij}, between any pair of participants i and j is the degree to which the standards of participants i and j are consistent with each other, and is defined quantitatively by the following two terms:

$$\begin{aligned} D_{ij} &\equiv x_i - x_j \\ U(D_{ij}) &\equiv U(x_i - x_{\mathrm{j}}) \end{aligned}$$

where $U(D_{ij})$ is the expanded uncertainty in D_{ij} at a 95 % level of confidence.

Note that while the definition of the unilateral DoEs involves the KCRV, this definition of the bilateral DoEs does not depend on the KCRV[1]. The value for the bilateral DoE is simply the difference between the values submitted by the two laboratories. However, the evaluation of the uncertainty, $U(x_i - x_j)$, is not as straightforward because it is not obtained by simply combining the individual participant standard uncertainties, $u(x_i)$ and $u(x_j)$, in quadrature and applying the coverage factor corresponding to a 95 % confidence level. Instead, depending on the statistical model adopted for the measured values and the topology of the KC (for example when x_i and x_j are performed in different loops), $U(D_{ij})$ can include additional uncertainty components besides the reported uncertainties for x_i and x_j, which prevents the presentation of a formula that is universally valid.

[1] The definition given here deviates from that given in [3], which, although given only in words, seems to be $D_{ij} \equiv D_i - D_j \equiv (x_i - x_{\mathrm{KCRV}}) - (x_j - x_{\mathrm{KCRV}})$, so involves the KCRV.

2.2 Key Comparison Reference Value in CCT Key Comparisons

The CIPM MRA documents do not give a definition of the KCRV, nor do they specify how the KCRV should be computed—they only deal with how the KCRV should be interpreted. In this section, we first discuss the meaning to be attributed to the KCRV in CCT KCs, then we give a definition of KCRV in CCT KCs. The subject of how the KCRV should be computed will be addressed in the remaining sections of this Guide.

2.2.1 The Meaning to be Attributed to the KCRV in CCT KCs

2.2.1.1 The meaning attributed to the KCRV in the CIPM MRA and in the literature

The CIPM MRA [3] states that:
(A) "*In most cases, the KCRV can be considered to be a close approximation to the SI value of the measurand"*.

The CIPM MRA-G11 document [4] states that:
(B) "*Each KCRV is considered to be a close approximation of the true value of the measurand*".

However, as acknowledged by the Technical Supplement to the Arrangement, some Consultative Committees (CCs) have difficulties in relating the KCRV to the SI value of the measurand. Some authors (see, for example, [5]) even question the possibility for any KC to quantify the consistency of a standard to the true value of the measurand, arguing that, in order to achieve such a goal, additional information and assumptions would be needed. Other authors state that obtaining both consistency with the true value of the measurand and mutual consistency among the participants in a KC is attainable in rather few cases [6].

Table 1: Statements on the meaning attributed to the KCRV in CCT-KCs (from the corresponding final reports).

KC	Ref.	Statements on the meaning attributed to the KCRV
CCT-K1	[7]	*No additional significance is attached to the KCRV with respect to 'best estimate' of T_{90}*
CCT-K2	[8]	*The KCRV represents a common baseline against which all laboratory values can be compared*
CCT-K3	[9]	The participants decided not to compute the KCRV
CCT-K4	[10]	*The value of the so-defined KCRV must not be related to the ITS-90 value of the corresponding fixed point. The KCRV has no physical meaning and is used only as a notational shorthand for presenting a common baseline against which all laboratory values can be compared*
CCT-K5	[11]	No statement on the meaning of the KCRV
CCT-K6	[12]	*The KCRV has no absolute significance outside the comparison*
CCT-K7	[13]	*It is clear that this KCRV is not the best possible approximation of the true SI value*
CCT-K8	[14]	*The KCRV given in this comparison has no absolute significance, that means it does not represent a reference value in the SI, only a comparison parameter*
CCT-K9	[15]	*The KCRV is not meant to have any intrinsic meaning – it is a computational tool allowing comparisons across labs linked through measurements at the pilot lab*
CCT-K10	[16]	No statement on the meaning of the KCRV
CCT-K7.2021	[17]	No statement on the meaning of the KCRV

2.2.1.2 The meaning attributed to the KCRV in the past CCT Key Comparisons

It appears that the large majority of the pilots and participants in the past CCT KCs shared the concerns expressed in [5] since, in nearly all past CCT KCs, the participants refrained from attributing the KCRV the meaning of best (or close) approximation to the SI value (or true value). In Table 1 we report the statements made by the CCT KC participants on the meaning of the calculated KCRV.

2.2.1.3 The meaning to be attributed to the KCRV in CCT Key Comparisons

In order to establish the meaning that can be reasonably attributed to the KCRV in CCT KCs, we analyze in detail the measurand in CCT thermometry and hygrometry KCs.

First of all, in all completed and running CCT KCs, the measured quantity (the measurand) is not an SI quantity, because the CCT KCs are based on the International Temperature Scale of 1990 (ITS-90) [18]. The ITS-90 is a practical temperature scale, that has been constructed in such a way that the quantity T_{90} that it identifies is a close approximation to another quantity T (thermodynamic temperature, which is the SI quantity) according to the best estimates at the time the scale was adopted (1990). However, since the adoption of the ITS-90, the substantial advancements in the measurement of T (primary thermometry) have evidenced significant differences between T and T_{90} [19], to the extent that nowadays T_{90} cannot be considered anymore a close approximation to T. Moreover, the same technological advancements in primary thermometry are allowing direct and practical measurements of T in some temperature ranges, and the CCT is considering starting to run KCs in terms of T in parallel to the traditional KCs (in terms of T_{90}).

In conclusion, in the past and current CCT KCs, the KCRV cannot be considered to be a close approximation to the SI value of the measurand (as done in [3]). Statement (B) in section 2.2.1.1, which is more general than statement (A) in the same section, and refers to the true value of the measurand (independently from it being an SI quantity or not), is regarded as more appropriate for CCT KCs. In the future, when the CCT will run KCs in terms of thermodynamic temperature (SI), the statement in [3] will be applicable to their measurand.

2.2.1.3.1 The meaning of the KCRV in CCT thermometry key comparisons

Let's consider a CCT KC of SPRTs calibrated at the ITS-90 fixed points (this is the case of CCT-K2, CCT-K3, CCT-K4, CCT-K7 and CCT-K9, so the majority of thermometry KCs). In such a KC, each participant realizes its local ITS-90 temperature scale, based on its local set of fixed-point cells, then calibrates its transfer artifact (SPRT) against its realized fixed-point temperatures, and finally transports the SPRT to the pilot's laboratory, where the pilot compares the travelling SPRT to its own fixed-point cells.

Strictly speaking, the measurand in such a KC is the resistance ratio, W, between the SPRT resistance at a given fixed-point (for example, the indium freezing point) and the SPRT resistance at the triple point of water. The ITS-90 definition assigned to the freezing point of indium a reference temperature $T_{90} = 429.7485$ K and a reference resistance ratio $W_{r,In} = 1.60980185$. This does not mean that the true value of the measurand in the KC is 1.60980185. The resistance ratio

of the specific SPRT used as transfer artifact has a true value which is different from the reference value.

In fact, for the purposes of the KC, the true value of the SPRT ratio, W, has very marginal relevance. In the KC, the SPRT is only a means to store the information about the local scale realizations of the participants. The participants measure different values, $W_{i,\mathrm{In}}$, of the resistance ratio at the indium freezing point not because of their different abilities in measuring electrical resistances, but (primarily) because they realize different freezing point temperatures of indium (because of different batches of high-purity materials, different thermal environments produced by different furnaces, …). The goal of the KC is to find out the differences (in temperature!) between the individual ITS-90 scale realizations of the participants. In fact, the KCRV and the DoEs are expressed in terms of temperature and not in terms of resistance ratio. From this point of view, it can be said that the measurand W is only a proxy of the real measurand, the temperature of the freezing point of indium, which is an unequivocally defined thermodynamic equilibrium state (freezing point of an ideally pure material), although a procedure-defined measurand (all participants are using the same method: fixed-point cells with thermometer wells accommodated in furnaces/cryostats/thermal baths).

Note that the value of this real measurand is assigned by the ITS-90 definition (for indium: $T_{90} = 429.7485$ K). All participants assign to the indium freezing point the same value of $T_{90} = 429.7485$ K. So, obviously, the goal of the participants is not measuring the T_{90} values at the fixed points. The goal of the participants is to compare the respective local realizations of the temperature of the indium freezing point, using as a proxy the measured resistance ratio.

What is the meaning of the KCRV in this case? Obviously, the KCRV is not an estimate of the SI value, because the ITS-90 temperature T_{90} is not an SI quantity. On the other hand, it is not correct to state that the KCRV has no physical meaning. Because the KCRV is obtained by pooling the individual independent results of the participants using an appropriate statistical procedure, the KCRV gives the statistically most likely position of the ideal fixed-point realization. The deviation from the KCRV of each participant ($D_i \equiv x_i - x_{\mathrm{KCRV}}$) also has a physical meaning: it tells the participant its deviation from the statistically most likely position of the ideal fixed-point transition. In conclusion, if we interpret the measurand (the resistance ratio) as a proxy for the ideal freezing point temperature of indium when measured with the procedure defined by the ITS-90 (fixed point cells, furnaces, SPRT, …), we can say that the KCRV can be considered to be related to an estimate of the true value of the measurand (though not SI), as in [3].

2.2.1.3.2 The meaning of the KCRV in CCT hygrometry key comparisons

A similar reasoning to that applied above to thermometry KCs can be also applied to CCT hygrometry KCs (CCT-K6 and CCT-K8). In hygrometry, similarly to thermometry, each participant realizes its local dew/frost-point ITS-90 temperature scale of humid gas (air or other gases), striving to realize a sample with a known thermodynamic state, which is convertible to the following thermodynamic equilibrium state: complete water saturation of gas (air) at a known pressure (e.g. atmospheric) and temperature.

The difference from a thermometry KC is that, in a thermometry KC, the measurand (temperature) is undergoing a first-order phase transition, and the phase transition acts as a temperature controller

at the temperature of the phase transition itself. In a dew/frost-point temperature KC, the measurand (dew/frost point temperature) is not undergoing a first-order phase transition (with associated latent heat). For this reason, while the measurand values in thermometry KCs are "forced by nature" at the phase transitions of the materials, in dew/frost point temperature KCs nominal measurand values are to be agreed in the protocol (for example, 30 °C, 50 °C, 65 °C, 80 °C, 85 °C, 90 °C and 95 °C in CCT-K8), and the participants are asked to realize temperatures within ±0.5 °C of the agreed nominal values.

The transfer artifact in dew/frost point temperature KCs is commonly a state-of-the-art chilled-mirror type of dew point hygrometer. The measurand, the resistance of a PRT embedded under the surface of the chilled-mirror, at the agreed nominal values of dew/frost point temperature, is transformed into a temperature value using the IEC 60751 standard. To cope with the fact that different participants, when measuring at the same nominal dew/frost point temperature, are in fact measuring at different temperatures (within ±0.5 °C), the difference Δt between the measured dew/frost point temperature (PRT resistance transformed into IEC 60751 temperature) t_d^{hyg} and the applied dew/frost point temperature (generated by the local primary dew-point generator) t_d^{ref}: $\Delta t = t_d^{\mathrm{hyg}} - t_d^{\mathrm{ref}}$ is formed. In this way, the fact that each participant measures the transfer artifact at slightly different dew/frost point temperatures becomes irrelevant, under the assumption that the behavior of the transfer artifact is the same within ±0.5 °C. Again, in the same way as in a thermometry KC, the transfer artifact is used to store the information on the individual local dew/frost-point temperature scales, and what is actually compared in the KC are the local dew/frost-point temperature scales of the participants. Although the participants are not realizing the same measurand, the use of a stable transfer artifact allows them to obtain an estimate of the differences between the realized primary dew-point temperature standards (the generators) for the same measurand (though nominal).

Again, although the numerical value of the KCRV has no meaning in terms of an SI value, the position of the KCRV represents the most likely position of the ideal dew/frost point realization, and the deviations from the KCRV ($D_i \equiv x_i - x_{\mathrm{KCRV}}$) are the deviations of the participants' realizations from the statistically most likely position of the ideal realization. In this sense, we can say that the KCRV is an estimate of the true value of the measurand, as in [3].

2.2.2 Definition of KCRV in CCT Key Comparisons

As shown in the previous sections, the quantity measured by the participants (which, strictly speaking, is the measurand) in CCT KCs is just a proxy of the quantity that the participants are interested in comparing. As a consequence, the KCRV in CCT KCs is expressed by a value that is not directly "*the best estimate that can be made of the true value of the measurand*". Rather, it is a value that is related to "*the best estimate that can be made of the true value of the measurand*".

For this reason, we define the KCRV in CCT KCs as:
a value that is related to the best estimate that can be made of the true value of the measurand from the KC measurements.

2.3 Consistency of the Results with the KCRV, with One Another and Mutual Consistency

Before discussing further, we clarify the exact meaning of different concepts of consistency used by the MRA in [3] and [4]: consistency of the results with the KCRV, consistency of a pair of results with one another, and mutual consistency of a set of results:

(A) *Consistency of a measurement result with the KCRV*:

A measurement result, x_i, is said to be consistent with the KCRV, x_{KCRV}, when the difference $x_i - x_{\mathrm{KCRV}}$ does not differ (statistically and substantively) significantly from zero. The statistical significance depends on the probability distribution of the difference. The substantive significance depends on a determination, made by the relevant Consultative Committee (CCT in our case) about how large the absolute value of that difference can be without it impacting the comparability of measurement results in practice.

(B) *Consistency of two measurement results with one another*:

Two measurement results x_i and x_j are said to be consistent with one another when the difference $x_i - x_j$ does not differ (statistically and substantively) significantly from zero. If there are N participants, then $N(N-1)/2$ such differences can be contemplated. Deciding whether they are significantly different from zero therefore involves $N(N-1)/2$ tests of the hypothesis of no difference. The conclusions of these tests should take into account the fact that the larger the number of tests, the larger the probability that at least one of the tests will suggest an incorrect conclusion.

(C) *Mutual consistency of a set of measurement results:*

A set of measurement results is said to be mutually consistent (with one another, regardless of what the KCRV or the true value of the measurand may be), when the dispersion of the measured values is not (statistically and substantively) larger than what the reported uncertainties suggest that it should be. Cochran's Q test [20] is often used to determine the statistical significance of any apparent mutual inconsistency.

While [3] gives the impression that, if goal (A) is satisfied (consistency of all results with the KCRV), then it follows that goal (B) (consistency between each pair of results) is also satisfied, this is not true. In general, given a set of results, there is no definite logical implication (for example, (A) ⇒ (B), or (A) ⇒ (C)). In fact, it is possible for all the individual results to be consistent with the KCRV, and at least one pair of participants to have inconsistent results (in this case (A) ⇒ (B) is not true). It is possible also to have cases in which (A) and (C) are true and (B) is not true for at least one pair.

Obviously, an ideal set of KC results would simultaneously satisfy all three types of consistency: consistency of all results with the KCRV, consistency of all pairs of results with one another, and consistency of the whole set of results. This is hardly ever achieved.

3. Principles to be Adopted and Technical Protocol

The design of the Technical Protocol is crucial for the success of a KC, because no analysis method can save poor data collection. Moreover, some decisions taken in the early phase of the life of a KC, namely when designing the Technical Protocol, have an impact on the later phase of the analysis of the measurement results. In this section we provide guidance on such decisions.

The CIPM-MRA-G11 document [4] states that the Technical Protocol, for CIPM KCs, should include a description of the method to be used to determine the KCRV. However, as clarified by Koepke *et al*. [21], the appropriate method for data reduction cannot be selected before examining the KC measurement results because the data reduction should be consistent with a statistical model that is demonstrably adequate for the measurement results (see principle 4 below). Therefore, the method to be used to determine the KCRV cannot be specified in the Technical Protocol in advance of obtaining and examining the measurement results from all the participants.

However, the Technical Protocol should at least set the general principles to be followed in the analysis of the results. The CCT endorses the 6 principles enunciated in [21], reformulated by Possolo *et al*. [22], and reproduced here, as follows:

1) Degrees of equivalence (DoEs) will be computed for all participants regardless of whether or not their results are included in the calculation of the KCRV and associated uncertainty, except for any participants that have chosen to withdraw from the KC. The selection of some results for the computation of the KCRV, and the exclusion of others, should be based on substantive reasons that are documented in the KC Final Report. The mere fact that a measured value lies far from the bulk of the others alone is insufficient reason to set it aside, even if a statistical test suggests that it is an "outlier" [23, 24].

2) No measured value should dominate the KCRV value "automatically," simply because the associated measurement uncertainty is much smaller than the uncertainties associated with the other measured values. To this end, procedures for data reduction should include some damping mechanism to limit the influence that measured values with unusually small associated uncertainties will have upon the KCRV. This provision is particularly influential when the measurement results are mutually inconsistent, and the uncertainties reported for some of them are much smaller than the uncertainties reported for the bulk of the others.

3) Measurement methods should be sufficiently well characterized to warrant confidence in the belief that the measured values, taken as a group, are roughly centered on the true value of the measurand. Participating laboratories should have previously demonstrated sufficient competence based on satisfactory performance in previous KCs, pilot studies, or proficiency tests. If all the measured values tend to be too low or too high relative to the true value of the measurand, no statistical procedure that relies on the data alone will be able to detect this and correct the KCRV accordingly, but the DoEs may, even in such cases, still be comparable and informative.

4) A model should be formulated that explicitly relates the measured values to the true value of the measurand. Furthermore, the estimation of the true value, and the evaluation of the associated uncertainty, should be consistent with the statistical model and with some principle of estimation whose general reliability is widely recognized. The calculation of the KCRV is contingent on such

a model and on the choice of the optimality criterion that the KCRV is intended to satisfy. Suppose that the model specifies that the measured values are equal to the measurand plus random, mutually independent, Gaussian measurement errors. This alone does not suffice to justify combining the measured values in a particular way (say, as a weighted average of the measured values). An additional criterion is needed: for example, one that seeks to minimize the mean squared error of the KCRV, or one that requires a KCRV with minimal absolute error. Therefore, the choices that one needs to make to determine the KCRV should take into account the purpose that the KC is intended to serve.

5) The statistical model underlying data reductions should be able to detect, evaluate and propagate dark uncertainty, which manifests itself as dispersion of the measured values in excess of what the reported uncertainties suggest that it should be, and thereby accounts for mutual inconsistency of the measurement results. Failure to recognize and propagate dark uncertainty generally yields DoEs with uncertainties that are too small. Both the average and the median of the measured values, with or without the conventional weights (which are proportional to the reciprocal of the squared reported uncertainties) ignore dark uncertainty, and thus tacitly assume that the results are mutually consistent. Furthermore, the reliability of the reported uncertainties, often expressed in terms of the effective numbers of degrees of freedom that support them, should be taken into account and hence should be reported routinely, as required by CIPM guidance for KCs.

6) Degrees of equivalence should be computed consistently with their primary goal of identifying participants with "unusual" results, in the sense that their measured values lie beyond the range allowed by the model. In particular, each DoE should include: i) the uncertainty reported for the corresponding measured value; ii) the uncertainty associated with the KCRV; iii) the correlation between the measured value and the KCRV; and iv) any dark uncertainty that has been detected during data reduction.

3.1 Measurement Schemes

Depending on the number of participants, the need to distribute the measurement workload among the pilot, copilots and participants, and the properties of the transfer artifact (transportability, stability), different measurement schemes (topologies) can be considered.

In the 'collapsed star' topology, each participant measures their own transfer artifact before transporting it to the pilot laboratory. The pilot then measures the transfer artifacts delivered by each participant. Finally, the transfer artifacts are returned and each participant remeasures its own transfer artifact to check for any drift during the comparison. The advantages of this topology are the reduction in the time taken to complete the comparison, since all the artifacts are intercompared at the pilot laboratory, and a potential reduction in drift occurring during transportation. The disadvantage is the considerable workload on the pilot. The collapsed star topology has been adopted in many CCT KCs: CCT-K1, CCT-K2, CCT-K3, CCT-K7, CCT-K9 and CCT-K7.2021 (see Appendix 1).

In the simplest 'round robin' topology, a single artifact, originating from the pilot laboratory, is measured by all participants. The transfer artifact is transported sequentially to each laboratory, who measures the artifact and sends it on to the next participant. Eventually, the transfer artifact returns to the pilot laboratory, who remeasures it to check for drift. The advantage of this topology

is the reduction of the workload on the pilot compared with the collapsed star topology. The disadvantage is the risk associated with the transport of the same transfer artifact to all the participants' laboratories: the transfer artifact could be damaged or drift during shipping or during its stay at the participants' laboratories. In order to mitigate this risk, modified versions of this topology involve either returning a single transfer artifact to the pilot laboratory one or more times along the circulation loop (allowing interim drift checks) or having two (or more) circulation loops each with its own transfer artifact (the latter was the case of CCT-K6 and CCT-K8).

Mixed versions of the collapsed star and round robin are also possible. For example, in CCT-K4, the pilot's transfer artifact was transported sequentially to three copilots, and each copilot had their own transfer artifact circulated to the participants in their respective loops.

4. Screening of the Initial Data Set

Upon completion of the KC measurements, and after receiving the measurement reports from all the participants, the pilot assembles the initial data set. For a KC with *N* participants, the initial data set is a set of *N* triplets (x_i, u_i, ν_i), where x_i is the result of participant *i*, u_i is the combined standard uncertainty of participant *i*, and ν_i is the number of degrees of freedom associated with u_i.

It is recommended that the degrees of freedom of each uncertainty component contributing to u_i also be reported, regardless of how they may be used in the analysis of the results (see section 6). Each participant should also classify the uncertainty components as either random or systematic, because this information is needed to properly take correlations into account.

The initial data set is inspected by the pilot using appropriate graphical methods (see CCQM Guidance Note [1]). At a minimum, the measured values x_i and their corresponding standard or expanded uncertainty bars should be depicted in a plot whose caption must state whether the uncertainty bars represent standard or expanded uncertainties, to identify:

- measured values x_i whose location appears inconsistent with the majority of the measured values;
- uncertainties that are unusually large or small;
- measured values x_i that deviate substantially from some preliminary estimate of the KCRV relative to their reported uncertainties.

Even without resorting to the use of outlier tests, such a simple visualization of the measurement results allows the identification of anomalous results and enables a visual assessment of the dispersion of the results. Figure 2, taken from CCQM Guidance Note [1], shows some of the main features of a metrology comparison data set. One of the 4 regions depicted in the figure will best represent the data set:

A = mutually consistent results (no excess variance, no outliers);

B = general over-dispersion or understatement of uncertainty affecting most of the participants;

C = generally consistent results with a small number of apparent outliers;

D = over-dispersion combined with some particularly extreme apparent outliers.

However, Figure 2 does not depict a general configuration that can be sometimes encountered in KCs: when the measured values have an asymmetric distribution, trailing either to the left, or to the right of the center.

Visually-identified outliers can be confirmed with statistical outlier tests, or tests based on non-parametric or robust statistics. A common non-parametric indicator of an outlier is a result outside the inter-quartile range $[Q_1, Q_3]$ that lies more than $1.5(Q_3 - Q_1)$ below Q_1 or above Q_3. This implies that a value sampled from a Gaussian (or normal) distribution will be falsely declared as an outlier with probability 0.7 %. Additional tests, such as the Cochran's Q test [20] and the (modified) Birge ratio test [25] can be considered.

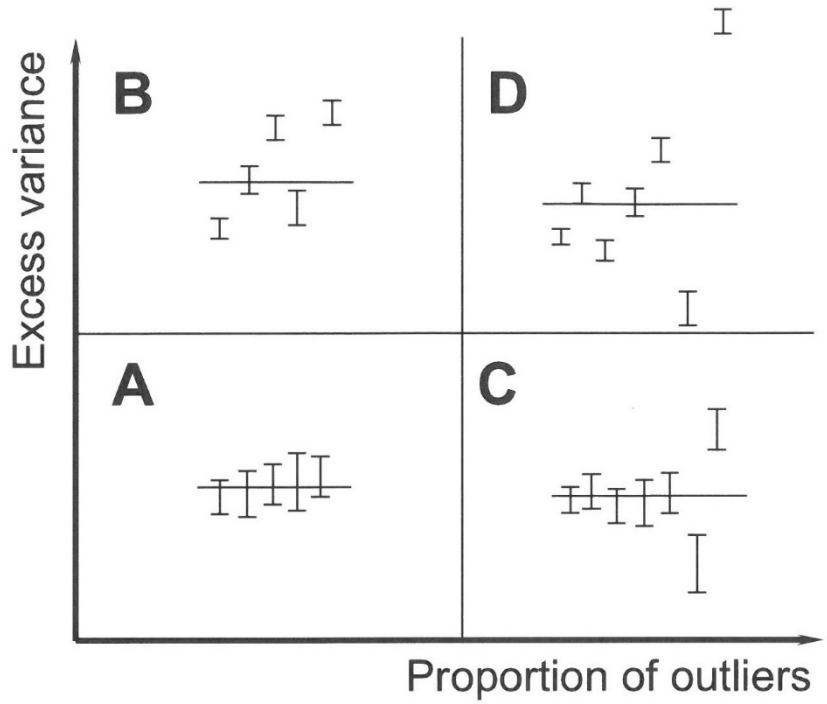


Figure 2: Features of importance in metrology comparison data. Vertical bars denote 95 % coverage intervals or expanded uncertainty intervals. A: mutually consistent data; B: evidence of general over-dispersion (or underestimation of uncertainty affecting most of the participants); C: generally consistent results with a small number of outlying values; D: general over-dispersion combined with some particularly extreme values. Taken from [1].

5. Dealing with Anomalous Results, Amendments and Withdrawal of Results

In case the initial screening of the measurement results using graphical methods and statistical tests of outliers identifies anomalous results, the pilot should follow up with the participant as detailed in [4]. When the participant shows evidence of arithmetic, typographical, or transcription error, the result can be corrected. When the participant cannot identify any such error, and there is no substantive reason to set the results aside (for example, incorrect calibration, notable deviation from the KC's technical protocol, etc.), then the participant's result stands and is not removed from the data set.

Statistical tests alone are not sufficient to justify removal of any measurement result from the data set used to calculate the KCRV. There must be a substantive reason for excluding any measurement result from the calculation of the KCRV. Laboratories excluded from the KCRV calculation will, however, be included in data summaries and plots, and DoEs will be calculated for them after calculation of the KCRV.

Participants are allowed to withdraw their results only if a reason not attributable to the performance of the laboratory can be identified (for example, if an excessive drift or a malfunction is detected in the transfer standard).

The initial data set, eventually modified by correcting arithmetic, typographical or transcription errors, and after excluding results based on substantive reasons, becomes the final data set. The final data set is used in the subsequent analysis (and in the determination of the KCRV).

6. Statistical Tests

As will become evident in the following sections, statistical tests performed on the final data set are a fundamental step in the selection of the procedure for calculating the KCRV. The goal of the statistical tests is to listen to the data set, and to coax it to tell you how it would like to be analyzed.

In this section we describe three statistical tests, which are those used in the NIST Decision Tree [21], as the results from these tests develop a recommendation for how the measurement results should be modeled and analyzed. We also explain the meaning to be attributed to the results of such tests. Obviously, the KC pilots and participants are free to apply other statistical tests to help them navigate the NIST Decision Tree, or to select a model and data reduction procedure that is not among those implemented in the NIST Decision Tree.

The NIST Decision Tree recommends that the following statistical tests be performed: a homogeneity test, which tests the mutual consistency of the measurement results using Cochran's Q test [20]; a normality (Gaussian shape) test, which tests whether the results are normally distributed using the Shapiro-Wilk test [26]; and a symmetry test, which tests whether the results are symmetrical using the Miao-Gel-Gastwirth test [27].

For each of these tests, the *p*-value is the probability, assuming that the hypothesis under test is true, of obtaining a result at least as extreme as that observed, owing to random sampling variation. A small *p*-value therefore provides evidence against the hypothesis under test.

For example, suppose the outcome of the homogeneity test is $p = 0.03$. If the measurement results are in fact drawn from a homogeneous population, there is a probability of only 3 % of obtaining a test result at least as extreme as that observed. The small *p*-value therefore provides evidence against the hypothesis of homogeneity.

The question then becomes: how small does the *p*-value need to be before the evidence is considered sufficient to reject the hypothesis under test? This is not determined by the *p*-value itself, and the NIST Decision Tree does not require a particular significance level or make rejection of a hypothesis compulsory. Rather, the *p*-value provides information that can be considered,

together with other statistical and scientific information, when deciding how the measurement results should be treated.

If a significance level, α, is chosen, then a rule such as rejecting the hypothesis when $p \leq \alpha$ can be applied. For example, with $\alpha = 0.05$, a p-value of 0.03 would lead to rejection of the hypothesis according to that rule. However, this does not mean that there is a 3 % probability that the decision to reject the hypothesis is wrong. The 3 % refers instead to the probability of obtaining a result at least as extreme as that observed, assuming that the hypothesis is true. A Type I error occurs when the hypothesis is rejected even though it is true. The significance level, α, specifies the maximum probability of making a Type I error under the hypothesis when the test is applied according to its specified decision rule. It should not be interpreted as the probability that a particular rejection of the hypothesis is wrong.

The user also needs to consider another feature of statistical hypothesis tests: the probability of failing to reject a hypothesis when it is false. This is known as a Type II error. The complement of this probability, namely the probability of rejecting a false hypothesis, is called the power of the test. The power of a statistical test generally increases with the size of the data set. It also depends on the extent and nature of the deviation of the data from what is described by the hypothesis under test. Thus, a failure to reject a hypothesis does not necessarily provide strong evidence that the hypothesis is true, particularly when the data set is small or the departure from the hypothesis is small.

The NIST Decision Tree was designed to provide substantial protection against the adverse consequences of incorrect decisions that can arise when interpreting the p-values from the three tests it uses. The NIST Decision Tree therefore recommends different thresholds for the three statistical tests:

- **Homogeneity test** (Cochran's Q test [20]): It is considered appropriate to reject the hypothesis of homogeneity even for relatively high p-values, because incorrectly assuming homogeneity can have significant consequences for the subsequent analysis. Accordingly, it is recommended that the hypothesis of homogeneity be rejected relatively readily, when the p-value is 10 % or smaller.
- **Normality test** (Shapiro-Wilk test [26]): The generally small number of participants in KCs calls for a moderately conservative stance regarding distributional shape. Accordingly, it is recommended that the hypothesis of Gaussian shape be rejected only when the p-value is smaller than 5 %.
- **Symmetry test** (Miao-Gel-Gastwirth test [27]): Since asymmetry can easily be confused with heavy tails in the probability distribution of laboratory or method effects, particularly in the small data sets commonly encountered in KCs, it is recommended that the hypothesis of symmetry be rejected only when the p-value is smaller than 1 %.

These thresholds should not be interpreted as probabilities that the respective hypotheses are true or false, nor as probabilities that a decision based on the test is wrong; they are decision criteria chosen to reflect the consequences of the different possible assumptions in the KC analysis. The NIST Decision Tree is effectively recommending different Type I error tolerances for different assumptions because the consequences of incorrectly retaining those assumptions are different.

7. Bilateral Degrees of Equivalence

The bilateral degree of equivalence (also known as pairwise equivalence), D_{ij}, between two participants is simply the difference between the values measured by any pair of participants, together with its associated expanded uncertainty for 95 % coverage, $U(D_{ij})$:

$$D_{ij} = x_i - x_j$$

$$U(D_{ij}) = U(x_i - x_j)$$

Once the final data set is fixed, the bilateral degrees of equivalence can be computed. The definition and computation of a KCRV is not necessary for calculating the bilateral degrees of equivalence.

The only difficulty is the evaluation of the standard uncertainty $u(D_{ij})$, because it is not simply given by the combination of the individual participant standard uncertainties, u_i and u_j, in quadrature.

Depending on the statistical model adopted for the measured values x_i and on the topology of the KC, $u(D_{ij})$ can involve other uncertainty components besides the reported uncertainties for x_i and x_j, which prevents presenting a universally valid formula. Also, it may not be possible to evaluate either $u(D_i)$ or $u(D_{ij})$ in a closed form, and a Monte Carlo based evaluation may be the best approach.

A comprehensive treatment of this problem is beyond the scope of this document, so here we provide only an example of a situation that is often encountered in CCT KCs. For this specific example, we provide both the formal approach (basically, the approach provided in the "Guide to the expression of uncertainty in measurement" [28] (section 7.3)), and the approach usually adopted by the CCT KC pilots (section 7.1 and 7.2). Note that the NIST Decision Tree computes unilateral, but not bilateral, degrees of equivalence. The NIST Consensus Builder (available at https://consensus.nist.gov) computes both, including a correct evaluation of the expanded uncertainties that are an integral part of the degrees of equivalence.

7.1 Usual CCT Approach

We consider a KC in which each participant calibrates a transfer standard SPRT against its reference temperature and delivers the transfer SPRT to the pilot. The result of participant *i* is the resistance ratio W_{il} of the transfer standard *l* with standard uncertainty u_{il} and degrees of freedom ν_{il}. Note that u_{il} includes not only the uncertainty in generating the temperature T_i, but also the uncertainty in measuring the resistance ratio W_{il}. In the same way, another participant *j* measures the resistance ratio W_{jm} of the transfer standard *m* when it is in thermal equilibrium with its reference temperature T_j, with standard uncertainty u_{jm} and degrees of freedom ν_{jm}.

The pilot *P* measures the resistance ratios W_{Pl} and W_{Pm} of transfer standards *l* and *m* belonging to participants *i* and *j*, respectively, when they are in thermal equilibrium with the pilot's reference

temperature T_P (to distinguish the two separate measurements, we denote T_P by T_{Pl} and T_{Pm}), with standard uncertainties u_{Pl} and u_{Pm} and degrees of freedom ν_{Pl} and ν_{Pm}, respectively.
Although this is not the most complex case encountered in CCT KCs, it shows that, when computing $u(D_{ij})$, correlations need to be considered. The history of CCT KCs reveals that the CCT community has demonstrated good competence in dealing with correlations.

The bilateral degrees of equivalence are obtained from the following equations (where dW_r/dT is the derivative of the SPRT reference function with respect to temperature, calculated at the reference temperature):

$$T_{il} - T_{Pl} = \frac{W_{il} - W_{Pl}}{dW_\mathrm{r}/dT} + \delta_{Sl}$$

$$u^2(T_{il} - T_{Pl}) = \frac{u_{il}^2 + u_{Pl}^2}{(dW_\mathrm{r}/dT)^2} + u_{Sl}^2$$

$$T_{jm} - T_{Pm} = \frac{W_{jm} - W_{Pm}}{dW_\mathrm{r}/dT} + \delta_{Sm}$$

$$u^2(T_{jm} - T_{Pm}) = \frac{u_{jm}^2 + u_{Pm}^2}{(dW_\mathrm{r}/dT)^2} + u_{Sm}^2$$

$$D_{ij} = (T_{il} - T_{Pl}) - (T_{jm} - T_{Pm}) = \left(\frac{W_{il} - W_{Pl}}{dW_\mathrm{r}/dT} + \delta_{Sl}\right) - \left(\frac{W_{jm} - W_{Pm}}{dW_\mathrm{r}/dT} + \delta_{Sm}\right)$$

$$u(D_{ij}) = \sqrt{\frac{u_{il}^2 + u_{jm}^2 + u_{Pl,\mathrm{r}}^2 + u_{Pm,\mathrm{r}}^2}{(dW_\mathrm{r}/dT)^2} + u_{Sl}^2 + u_{Sm}^2}$$

$$\nu_{ij} = [u(D_{ij})]^4 \Big/ \left[\left(\frac{u_{il}^4}{\nu_{il}} + \frac{u_{jm}^4}{\nu_{jm}} + \frac{u_{Pl,\mathrm{r}}^4}{\nu_{Pl,\mathrm{r}}} + \frac{u_{Pm,\mathrm{r}}^4}{\nu_{Pm,\mathrm{r}}}\right)/(dW_\mathrm{r}/dT)^4 + \frac{u_{Sl}^4}{\nu_{Sl}} + \frac{u_{Sm}^4}{\nu_{Sm}}\right]$$

$$U(D_{ij}) = k_{ij} u(D_{ij})$$

In the model equation for D_{ij}, the terms δ_{Sl} and δ_{Sm} are included to take into account the drift of the transfer standards *l* and *m*, respectively, over the period of the KC. The best estimates for δ_{Sl} and δ_{Sm} can be taken as zero, but their corresponding uncertainties, u_{Sl} and u_{Sm}, and numbers of degrees of freedom, ν_{Sl} and ν_{Sm}, must be included when calculating the uncertainty in D_{ij}.

In the expression for $u(D_{ij})$, $u_{Pl,\mathrm{r}}$ and $u_{Pm,\mathrm{r}}$ are the standard uncertainties due to the pilot's random errors when measuring transfer standards *l* and *m*, respectively. The reason for using only the random uncertainties of the pilot in this expression is because the pilot measurements T_{Pl} and T_{Pm} are correlated due to the shared systematic effects between the two measurements. The

uncertainties arising from the pilot's systematic errors are not included because it is assumed that they are fully correlated between the two measurements and therefore cancel when subtracted.

The effective number of degrees of freedom of $u(D_{ij})$, ν_{ij}, is obtained by applying the Welch–Satterthwaite formula to the degrees of freedom ν_{il}, ν_{jm}, $\nu_{Pl,\mathrm{r}}$, $\nu_{Pm,\mathrm{r}}$, ν_{Sl} and ν_{Sm} of the corresponding standard uncertainties u_{il}, u_{jm}, $u_{Pl,\mathrm{r}}$, $u_{Pm,\mathrm{r}}$, u_{Sl} and u_{Sm}, respectively.

Finally, in the expression for $U(D_{ij})$, the coverage factor k_{ij} is obtained from the Student's *t*-distribution, $t_{0.95}(\nu_{ij})$, with the appropriate degrees of freedom ν_{ij}.

7.2 Additional Remarks for Humidity to the Usual CCT Approach

As described in section 2.2.1.3. the KCs in the field of humidity are in many aspects similar to the KCs in thermometry. One difference is that the participants traditionally repeat the measurement four times at each nominal dew/frost-point temperature. These four measurements must be reduced to a single aggregated result in order to calculate the KCRV. For the aggregation of the results, procedures similar to those advised in this document should be applied and the correlation between the repeat measurements should be taken into account. Another peculiarity is that, in the most recent humidity KCs, topologies with more than one loop and more than one artifact were used, resulting in more than one option to link measurement results from different loops. For these additional complications, solutions were worked out (see the CCT-K8 final report and Appendix 1).

7.3 Formal Approach

The bilateral degrees of equivalence, D_{ij}, are obtained from the four temperatures T_{il}, T_{Pl}, T_{jm}, and T_{Pm}. To simplify the notation, we denote these variables by x_1, x_2, x_3 and x_4, respectively. D_{ij} is a function of x_1, x_2, x_3 and x_4:

$$D_{ij} = f(x_1, x_2, x_3, x_4) = x_1 - x_2 - x_3 + x_4$$

with correlation between the variables x_2 and x_4.

When the input quantities are correlated, the appropriate general expression for the combined uncertainty $u^2(D_{ij})$ is:

$$u^2(D_{ij}) = \sum_{i=1}^{4} \left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i) + 2\sum_{i=1}^{3}\sum_{j=i+1}^{4} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j} u(x_i, x_j)$$

where $u(x_i, x_j)$ is the estimated covariance associated with x_i and x_j.

Note that the first term of this expression is the usual term encountered when the input quantities are uncorrelated, and the second term arises when correlations among the input quantities are present.

In our specific example, there is correlation only between the input quantities x_2 and x_4. In this case the general expression above reduces to:

$$u^2(D_{ij}) = \sum_{i=1}^{4} \left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i) + 2\frac{\partial f}{\partial x_2}\frac{\partial f}{\partial x_4}u(x_2, x_4)$$

Moreover, $\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial x_4} = 1$ and $\frac{\partial f}{\partial x_2} = \frac{\partial f}{\partial x_3} = -1$, so the expression reduces further to:

$$u^2(D_{ij}) = \sum_{i=1}^{4} u^2(x_i) - 2u(x_2, x_4)$$

We only need to estimate the covariance associated with x_2 and x_4. The covariance $u(x_2, x_4)$ is given by:

$$u(x_2, x_4) = \frac{u(x_2)u(x_4)}{\sqrt{1 + \frac{u_\mathrm{r}^2(x_2)}{u_\mathrm{s}^2(x_2)}}\sqrt{1 + \frac{u_\mathrm{r}^2(x_4)}{u_\mathrm{s}^2(x_4)}}}$$

where $u_\mathrm{r}(x_2)$ is the combined random uncertainty components of x_2, $u_\mathrm{s}(x_2)$ is the combined systematic uncertain components of x_2, and similarly for x_4.

8. Statistical Models

Four kinds of statistical models are commonly used for meta-analysis and in interlaboratory studies, which can be fitted to sets of measurement results using either classical or Bayesian versions: common mean [29], common mean with inflated uncertainties [24], fixed effects [30], and random effects [21].

Considering the main features of CCT KC measurands (see discussion in Section 2), random effects models generally tend to be the most appropriate models for the measurement results of CCT KCs. This class of models comprises a great variety of specific models suitable for data sets like those illustrated in the four panels of Figure 2, as well as data sets exhibiting other patterns. These models can be used successfully even when all the results depicted should be included in the calculation of the KCRV. In this document we focus on random effects models, and on techniques for estimating the parameters of these models, one of them being the KCRV.

A clarification is required for the meaning of the term “random effects”. The effects in question describe whether participants tend to measure similar measurands high or low persistently. That is, these effects can be interpreted as laboratory biases. Modeling these effects as random variables (this is the reason they are called “random”) does not really mean that they are variable: it means simply, and pragmatically, that their particular values are unknown to some extent, and that

probability distributions are used to characterize this state of knowledge. In other words, the randomness here is epistemic, not stochastic.

As clarified in Section 3 (principle 5), the statistical model can detect dark uncertainty, which is evaluated as a variance in excess of that expected from the reported uncertainties. When the presence of dark uncertainty is observed, the CCT must decide if it should affect the DoEs. We believe the answer depends on the purpose that the DoEs serve. If the DoEs serve to express each participant's subjective opinion on its capability, then they could exclude the uncertainty contribution from dark uncertainty. If the DoEs are intended to predict a future measurement result (either of one of the participants, or of a laboratory that did not participate in the KC but whose capabilities are judged to be similar to the participants), then the DoEs should recognize the dark uncertainty.

The random effects model is expressed by the model equation:

$$x_i = \mu + \lambda_i + \varepsilon_i$$

where x_i is the value measured by participant i, μ is the true value of the measurand, λ_i is the unknown laboratory effect of participant i, and ε_i is the laboratory-specific measurement error of participant i.

The λ_i are modeled as non-observable outcomes of independent random variables, all coming from the same distribution with mean 0 and standard deviation τ: in some versions of the model their probability distribution is Gaussian, in others it can be Laplace (or double exponential), or some other distribution. Most often, and in the absence of compelling evidence to the contrary, the measurement errors ε_i are considered to be values of independent Gaussian random variables with mean 0 and standard deviations u_i. These assumptions imply that the expected value (true value) of the measurand is μ. This reflects the belief that the participants' measurements are centered on the true value of the measurand, even if some participants tend to measure high and others low (as described by the λ_i).

The λ_i can be distinguished from the ε_i because the reported uncertainties $u_i = u(x_i)$ are an integral part of the measurement data. If the x_i are more dispersed than the u_i suggest that they should be, then the λ_i cannot all be zero, and a dark uncertainty τ is present. When the λ_i do not differ significantly from 0, there is no dark uncertainty ($\tau = 0$).

The model equation above is valid for the simplest comparison topology in which the same artifact is measured by all participants only one time. For the most general comparison topology, in which multiple artifacts are measured repeatedly by the participants, the model equation becomes [7]:

$$x_{i,j,k} = \mu_j + \lambda_i + \varepsilon_{i,j,k}$$

where the indices i, j and k identify the participant, the artifact and the index of a participant's measurement, respectively; μ_j is the true value of the j-th artifact; λ_i is each participant's bias; $\varepsilon_{i,j,k}$ is the random error associated with each measurement and is distributed about zero with a standard deviation s_{ijk}.

Thus, the variance in a participant's single measurement of the artifact value is:

$$\mathrm{var}\left(x_{i,j,k}\right) = \tau^2 + s_{ijk}^2.$$

A method for estimating τ is given in Section 9.4 below.

9. Statistical Procedures for Computing the Key Comparison Reference Value

Many different procedures have been used historically to compute the KCRV. The statistical model for the measured values alone does not stipulate the procedure that should be used to compute the KCRV, or the procedure to compute the DoEs. The optimal procedure depends both on the model and on the meaning of "optimal". In this section, we will review several of these procedures, and will explain the conditions under which they may be optimal. The procedures included here have been selected based on the availability of tools to implement them and are a small subset of available procedures for comparison analysis. The CCT does not require these specific methods to be used and pilots are free to use any method that they see fit. Refer to Borenstein *et al.* [29], Hartung *et al.* [31], Koepke *et al.* [21] and Possolo *et al.* [22], for comprehensive reviews of estimation procedures for different kinds of models.

9.1 Simple Average

The KCRV value x_{KCRV} and its standard uncertainty u_{KCRV} are given by:

$$x_{\mathrm{KCRV}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$u_{\mathrm{KCRV}}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - x_{\mathrm{KCRV}})^2$$

The simple average will be optimal (minimum mean squared error) when the measured values are sampled from a single Gaussian distribution. In particular, there is no dark uncertainty ($\tau = 0$) and all the x_i are assumed to have the same standard uncertainty, but the reported standard uncertainties are tacitly assumed to be unreliable and all are disregarded.

9.2 Uncertainty-Weighted Average

The KCRV value x_{KCRV} and its standard uncertainty u_{KCRV} are given by:

$$x_{\mathrm{KCRV}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$w_i = 1/u_i^2$$

$$\frac{1}{u_{\mathrm{KCRV}}^2} = \sum_{i=1}^{n} \frac{1}{u_i^2}$$

The uncertainty-weighted average will be optimal (minimum mean squared error) when the measured values are outcomes of Gaussian random variables with the same mean but possibly different standard deviations that are reliably estimated by the reported uncertainties. In particular, there is no dark uncertainty ($\tau = 0$). The formula for u_{KCRV} also assumes that each of the reported uncertainties is based on infinitely many (or, in practice, a large number of) degrees of freedom.

9.3 Median

The KCRV value x_{KCRV} and its standard uncertainty u_{KCRV} are given by (when the participant values x_i are arranged in increasing order):

$$x_{\mathrm{KCRV}} = \begin{cases} \frac{1}{2}\left(x_{n/2} + x_{(n/2)+1}\right), & n \text{ even} \\ x_{(n+1)/2}, & n \text{ odd} \end{cases}$$

$$u_{\mathrm{KCRV}}^2 = \frac{\pi}{2n}\hat{\sigma}^2$$

$$\hat{\sigma} = 1.483 \cdot \mathrm{MAD}$$

where MAD is the median of the absolute deviations: $\mathrm{MAD} = \mathrm{median}(|x_i - x_{\mathrm{KCRV}}|)$.

However, the formula given for the standard uncertainty $\hat{\sigma}$ can often be incorrect, for two reasons: i) the measured values may have rather different standard uncertainties, hence a weighted median in this case will be more appropriate than the simple median; ii) the formula strictly only applies to the median of large samples drawn from a Gaussian distribution. If this is the case, then there is no reason to use the median to begin with, and the simple average or uncertainty-weighted average should be used instead.

The median will be optimal in the sense of minimum mean absolute error when the measured values are sampled from a single Laplace distribution. In particular, there is no dark uncertainty ($\tau = 0$) and all the x_i are assumed to have the same standard uncertainty, but the reported standard uncertainties are tacitly assumed to be unreliable and are all disregarded. If σ is the true, common standard deviation of the Laplace measurement errors, then the standard uncertainty of the median is $\sigma/\sqrt{2n}$, for large n.

In general, the formula for the standard uncertainty of the median involves the value that the probability density of the measured values takes at the median of their common distribution, and it provides no more than an approximation whose accuracy increases with the number, n, of participants.

Considering that the number of participants in a KC is generally small, a better alternative is to evaluate the standard uncertainty of the median as half the length of a 68 % confidence interval

for the median by analogy with the Gaussian case (Possolo and Bodnar [32], which can be computed in fair generality as described by Hettmansperger and Sheather [33]).

9.4 DerSimonian-Laird Procedure (Adaptive Weighted Average, AWA)

The procedure suggested by DerSimonian and Laird [34] is suitable to fit the random effects model to the results from KCs both when $\tau > 0$ and when $\tau = 0$, and otherwise does not make specific distributional assumptions about either the λ_i or the ε_i introduced in Section 8.1. When $\tau = 0$, the procedure reduces to the conventional weighted average discussed in Section 9.2.

The KCRV value is a weighted average of the values x_i measured by the participants, with weights $w_i = 1/(u_i^2 + \tau^2)$ (note the difference from the weights used in 9.2):

$$x_{\mathrm{KCRV}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$w_i = 1/(u_i^2 + \tau^2)$$

An estimate of τ is obtained from:

$$\tau = \max\{0, \hat{\tau}_M\}$$

$$\hat{\tau}_M^2 = \frac{(Q - n + 1)}{(\sum_{i=1}^{n} u_i^{-2} - \sum_{i=1}^{n} u_i^{-4} / \sum_{i=1}^{n} u_i^{-2})}$$

$$Q = \sum_{i=1}^{n} u_i^{-2} \left(x_j - x_{\mathrm{KCRV}}\right)^2$$

The standard uncertainty of the KCRV is given by:

$$u_{\mathrm{KCRV}} = \frac{1}{\sqrt{\sum_{i=1}^{n} w_i}}$$

Neither this uncertainty evaluation for the KCRV, nor the uncertainty evaluations for the DoEs, take into account the fact that the estimate of τ above is based on the dispersion of the measured values, hence typically is supported by a fairly small number of degrees of freedom. The Adaptive Weighted Average implemented by the NIST Decision Tree [22] is a generalization of the DerSimonian-Laird procedure that addresses this shortcoming.

A similar procedure implementing the random effects model was proposed by Mandel and Paule [35]. In this procedure, the consistency of the data is tested using the Cochran's Q test. If the data is considered to be inconsistent, an iterative process is used to find a value of τ that gives consistent results.

9.5 Maximum Likelihood

This procedure is described by White [36] and by Koepke *et al*. [21] (subsection 5.1) in the metrological context, and in general by Searle *et al*. [37]. In this method, the maximum likelihood estimate of the KCRV is obtained by maximizing the likelihood function that corresponds to a random effects model with Gaussian λ_i and ε_i, with respect to two parameters: the overall mean (μ in the model equation), and the dark uncertainty (the standard deviation τ of the laboratory effects in the model equation), assuming that the reported uncertainties are based on very large numbers of degrees of freedom. An approximation to the standard uncertainty associated with the KCRV is obtained as a side-effect, so that the larger the number of participants, the better the quality of the approximation.

R function 'rma', defined in package 'metafor' [38], can fit both the random effects and fixed effects variants of the model via maximum likelihood estimation (MLE). The random effects version of the model can also be fitted (again using function 'rma') via restricted maximum likelihood (REML) estimation, which is generally preferable to MLE, especially when the number of participating laboratories is small [39].

9.6 Linear Pool

The linear pool is described by Koepke et al. [21]. It consists of representing each measurement result (measured value x_i and associated uncertainty u_i) by a probability density function $f_i(x)$ (Gaussian or some other), and computing a weighted average of these functions:

$$g(x) = \sum_{i=1}^{n} w_i \, f_i(x) / \sum_{i=1}^{n} w_i .$$

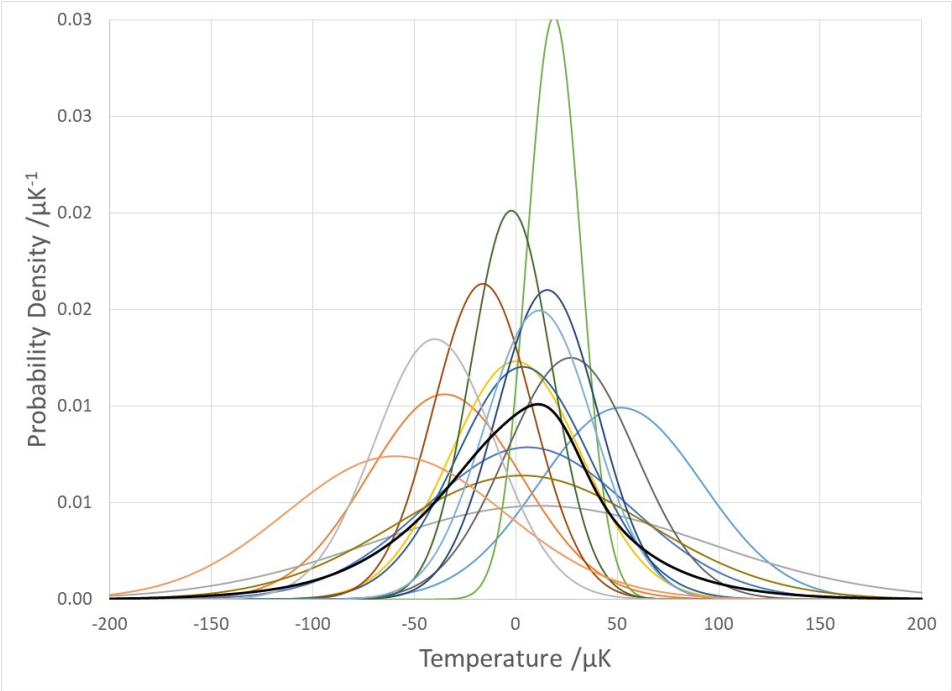


Figure 3: CCT-K7.2021 Key Comparison of triple-point-of-water cells: individual participants' distributions (thin curves) and linear pool distribution of results (thick black curve).

The KCRV is then obtained as the expected value of the distribution with probability density $g(x)$, which can easily be computed by Monte Carlo methods, and the standard uncertainty associated with the KCRV is the standard deviation of the same distribution.

Figure 3 shows the $f_i(x)$ for CCT-K7.2021, chosen to be Gaussian with means equal to the values measured by the participants, and with standard deviations equal to their reported uncertainties. The linear pool has a mean –0.4 μK and standard deviation 41.4 μK.

9.7 The NIST Decision Tree

The NIST Decision Tree (NDT) [2] is a useful tool that can be used in the analysis of KC results. The NDT is a web-based application that guides the user through a series of questions whose answers are informed by the results of statistical tests, to help the user decide on an appropriate statistical model and a data reduction procedure for the particular data set input by the user. It works as follows: 1) the user inputs the data set (the measured value and corresponding uncertainty reported of each participant) into a spreadsheet of the NDT; 2) the NDT performs statistical tests (homogeneity, symmetry and normality) on the data set; 3) the user, based on the results of the statistical tests, on the recommendation that the NDT makes, and on their scientific knowledge and judgement, selects the appropriate statistical procedure; and 4) the NDT carries out the analysis for the selected statistical procedure, and displays the results of the data reductions, including the KCRV and the degrees of equivalence, in the form of plots, tables and a downloadable PDF report.

A flow diagram of the NDT is shown in Figure 4. The NDT is traversed by answering a question at each node (orange), and moving forward according to the answer given to the question: if the answer is yes, you move to the left (green branch), if the answer is no, you move to the right (red branch), until a leaf (blue) is reached, which is the procedure recommended by the NDT for the input data set.

The NDT has been successfully used in the analysis of CCT-K7.2021 [17] and is a tool that pilots could use to aid their analysis of future CCT KCs.

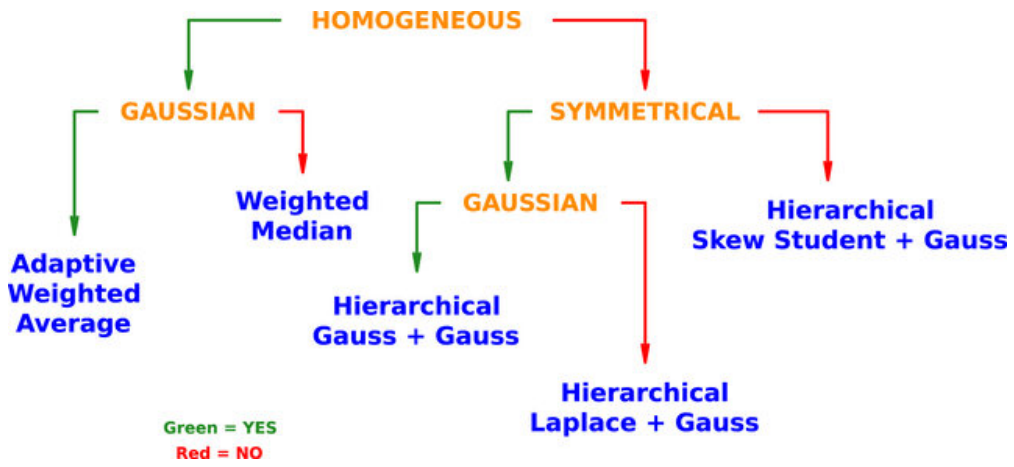


Figure 4: Flow diagram of the NIST Decision Tree.

10. Selection of the Statistical Procedure to be Used for Computing the Key Comparison Reference Value

Beyond the CIPM general statement about the KCRV given in [3, 4], no guidance is given about how to determine the KCRV of a KC.

The selection of the statistical procedure for calculating the KCRV should be based on:

1. Visual (graphical) inspection of the final data set (see section 4);
2. The outcomes of the applied statistical tests (see section 6);
3. The selected statistical model for the KC measurement (see section 8);
4. The scientific knowledge and judgement of the participants.

1 and 2 allow the pilot to establish which of the 4 regions depicted in Figure 2 best represents the final data set. Based on the past history, the results of CCT KCs are all represented by one of the following three regions (see Table 2):

- Region A: mutually consistent results with no overdispersion and no outliers.
- Border region between A and B: some overdispersion (small excess variance) and no outliers.
- Region C: general consistency of results with a small number of outliers.

Results represented by region D have not occurred in previous CCT KCs, so this case will not be considered.

Moreover, as discussed in section 8, the most appropriate statistical model for most CCT KCs is the random effects model, which includes a laboratory bias λ_i with a non-zero standard deviation τ (dark uncertainty), which is evidence of some over-dispersion.

Table 2: Regions of Figure 2 representing the results of each CCT key comparison (second column) and the corresponding adopted statistical procedure for calculating the KCRV (third column).

Key Comparison	Region of Figure 2	Procedure used for computing the KCRV
CCT-K1	Border region between A and B	Weighted average
CCT-K2	Region C	Weighted average
CCT-K3	Border region between A and B	Not computed
CCT-K4	Border region between A and B	Weighted average
CCT-K5	Region C	Median
CCT-K6	Border region between A and B	Weighted average
CCT-K7	Border region between A and B	Simple average
CCT-K8	Region A	Weighted average
CCT-K9	Border region between A and B	Simple average
CCT-K10	Region C	Median
CCT-K7.2021	Border region between A and B	DerSimonian-Laird

10.1 Results Represented by Region A

When the results are mutually consistent (region A, no outliers and no overdispersion), the uncertainty-weighted average (section 9.2) is the appropriate method to be used. However, a procedure such as the DerSimonian-Laird or Mandel-Paule procedure (adaptative weighted average, section 9.4) can also be used. The standard deviation τ will simply be close to zero and the KCRV value will coincide with the KCRV resulting from the application of the ordinary uncertainty-weighted average.

10.2 Results Represented by the Border between Region A and Region B

When the results are generally consistent but some overdispersion is present (evidenced by $\tau \neq 0$), a random effects model, such as the DerSimonian-Laird or Mandel-Paule procedure (section 9.4) is the appropriate method to be used. The uncertainty-weighted average should not be used when there is appreciable risk of significant over-dispersion. Using a model such as the DerSimonian-Laird or Mandel-Paule procedure prevents small participant-specific uncertainties from having an excessive influence on the KCRV.

10.3 Results Represented by Region C

When the results are generally consistent with one or more outliers present, an estimation procedure that is robust (resistant) to the presence of outliers should be used. The median is a simple robust estimator (section 9.3), though there are more efficient robust estimators, for example the M- and MM-estimators [40].

11. Conclusions

The main goal of this document is to provide guidance to CCT KC pilots and participants in the analysis of the results of CCT key comparisons. Although the analysis of KC results is traditionally addressed in the last stage of a KC lifetime, we recommend that CCT KC pilots and participants consult this document in the earliest phase of a KC, ideally when preparing the KC technical protocol.

There is no magic analysis method that is suitable for every possible KC data set (as there is no magic analysis method that can save poor measurement results). The KC pilots and participants, based on the results of the applied statistical tests and their unique scientific knowledge of the measurement, should select the most appropriate statistical model to be applied to the measurement. The reduction method (for the calculation of the KCRV and the unilateral degrees of equivalence) will follow naturally.

In spite of the unavoidable degree of discretion in the selection of the analysis method, this document has made an effort to harmonize all aspects of CCT KC analysis, trying to achieve, at least within the CCT community, a common understanding of the meaning to be attributed to the KCRV, and a common agreement on the general principles to be adopted when analyzing KC results.

Based on the past CCT KCs, we described only the analysis methods that are relevant for the CCT community. Obviously, by doing so, we limited the validity of this document to only CCT KCs.

Acknowledgements

The members of the CCT Working Group for Key Comparisons would like to express their deep gratitude to Dr Antonio Possolo for his invaluable support in the preparation of the document, and to Dr Ellie Molloy, Dr Peter Saunders and Dr Regina Derschermeir for their precious comments on the document.

References

[1] CIPM Consultative Committee for Amount of Substance, "CCQM Guidance note: Estimation of a consensus KCRV and associated degrees of equivalence," version 10, 2013.

[2] D Newton, A Koepke, A Possolo, M Winchester, "NIST Decision Tree User's Manual," 3rd edition, National Institute of Standards and Technology, 2024.

[3] CIPM, Technical supplement revised in October 2003. In *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates issued by National Metrology Institutes*; BIPM: Sèvres, France 2003.

[4] CIPM, MRA-G11, version 1.2. "Measurement Comparisons in the CIPM MRA: Guidelines for organizing, participating and reporting," 2025.

[5] C Elster, B Toman, "Analysis of key comparison data: Critical assessment of elements of current practice with suggested improvements," *Metrologia* **50** 549–555, 2013.

[6] U Panne, W Bremser, "Considerations with regard to the calculation of key comparison reference values and uncertainties," *Accreditation and Quality Assurance* **14** 281–287, 2009.

[7] R Rusby, D Head, C Meyer, W Tew, O Tamura, K D Hill, M de Groot, A Storm, A Peruzzi, B Fellmuth, "Final report on CCT-K1: Realizations of the ITS-90, 0.65 K to 24.5561 K, using rhodium-iron resistance thermometers," *Metrologia* **43** 03002, 2006.

[8] A G Steele, B Fellmuth, D I Head, Y Hermier, K H Kang, P P M Steur, W L Tew, "CCT-K2: key comparison of capsule-type standard platinum resistance thermometers from 13.8 K to 273.16 K," *Metrologia* **39** 551, 2002.

[9] B W Mangum, G F Strouse, W F Guthrie, R Pello, M Stock, E Renaot, Y Hermier, G Bonnier, P Marcarino, K S Gam, K H Khang, Y-G Kim, J V Nicholas, D R White, T D Dransfield, Y Duan, Y Qu, J Connolly, R L Rusby, J Gray, G J Sutton, D I Head, K D Hill, A Steele, K Nara, E Tegeler, U Noatsch, D Heyer, B Fellmuth, B Thiele-Krivoj, S Duris, A I Pokhodun, N P Moiseeva, A G Ivanova, M J de Groot, J F Dubbeldam, "Summary of comparison of realizations of the ITS-90 over the range 83.8058 K to 933.473 K: CCT key comparison CCT-K3," *Metrologia* **39** 179, 2002.

[10] H G Nubbemeyer, J Fischer, "Final report on key comparison CCT-K4 of local realizations of aluminium and silver freezing-point temperatures," *Metrologia* **39** 03001, 2002.

[11] M J de Groot, E W M van der Ham, R Monshouwer, "CCT-K5: Comparison of local realizations of the ITS-90 between the silver point and 1700 °C using vacuum tungsten strip lamps as transfer standards," 2008.

[12] S Bell, M Stevens, H Abe, R Benyon, R Bosma, V Fernicola, M Heinonen, P Huang, H Kitano, Z Li, J Nielsen, N Ochi, O A Podmurnaya, G Scace, D Smorgon, T Vicente, A F

Vinge, L Wang, H Yi, “Final report to the CCT on key comparison CCT-K6 – Comparison of local realisations of dew-point temperature scales in the range –50 °C to +20 °C,” *Metrologia* **55** 03005, 2015.

[13] M Stock, S Solve, D Del Campo, V Chimenti, E Mendez-Lango, H Liedberg, P P M Steur, P Marcarino, R Dematteis, E Filipe, I Lobo, K H Kang, K S Gam, Y-G Kim, E Renaot, G Bonnier, M Valin, D R White, T D Dransfield, Y Duan, Y Xiaoke, G Strouse, M Ballico, D Sukkar, M Arai, A Mans, M de Groot, O Kerkhof, R Rusby, J Gray, D Head, K Hill, E Tegeler, U Noatsch, S Duris, H Y Kho, S Ugur, A Pokhodun, S F Gerasimov, “Final report on CCT-K7: Key comparison of water triple point cells,” *Metrologia* **43** 03001, 2006.

[14] J G Gallegos, H Mitter, H Abe, S Bell, R Benyon, P Carroll, B I Choi, R Derschermeir, V Ebert, V Fernicola, L Wang, C Meyer, D Smorgon, T Vicente, “Final Report to the CCT on key comparison CCT-K8 – Comparison of local realisations of dew-point temperature of humid gas in the range 30 °C to 95 °C,” *Metrologia* **62** 03001, 2025.

[15] T Herman, M Chojnacky, K Hill, S Rudtsch, I Yang, P P M Steur, R Dematteis, G Lopardo, F Sparasci, C Martin, L Risegari, J Widiatmo, T Nakano, I Saito, K Quelhas, P Giorgio, J Sun, J Zhang, J Pearce, J Gray, “CCT-K9 Final Report – ITS-90 SPRT calibration from the Ar TP to the Zn FP,” *Metrologia* **60** 03001, 2023.

[16] H McEvoy, D Lowe, G Machin, M J Martin, K Anhalt, J Hollandt, M Sadli, F Bourson, Y Yamada, N Sasajma, X Lu, Y Yoo, H Yoon, C Gibson, A Todd, D Woods, “Final report for CCT comparison of ITS-90 realisations above the silver point using two transfer radiation thermometers and a set of high temperature fixed-point blackbody cells,” *Metrologia* **61** 03003, 2024.

[17] A Peruzzi, S Dedyulin, M Levesque, D Del Campo, B C Garcia Izquierdo, M E Gomez, K N Quelhas, M A P Neto, B M Lozano, L Eusebio, I Yang, F Sparasci, A Martin, L Risegari, P Saunders, E Molloy, X K Yan, J P Sun, X J Feng, J T Zhang, M-K Ho, T Nakano, J V Widiatmo, I Saito, E Ejigu, J Pearce, S Rudtsch, L Buenger, M Kalemci, A Uytun, C Bruin-Barendregt, M Panman, D R White, A Possolo, “CCT-K7.2021: CIPM key comparison of water-triple-point cells,” *Metrologia* **60** 03002, 2023.

[18] H Preston-Thomas, “The international temperature scale of 1990 (ITS-90),” *Metrologia* **27** 107, 1990.

[19] C Gaiser, B Fellmuth, R M Gavioso, M Kalemci, V Kytin, T Nakano, A Pokhodun, P M C Rourke, R Rusby, F Sparasci, P P M Steur, W L Tew, R Underwood, D R White, I Yang, J Zhang, “Update for the differences between thermodynamic temperature and ITS-90 below 335 K,” *Journal of Physical and Chemical Reference Data* **51** 043105 2022.

[20] W G Cochran, “The combination of estimates from different experiments,” *Biometrics* **10** (1):101–129, 1954.

[21] A Koepke, T Lafarge, A Possolo, B Toman, “Consensus building for interlaboratory studies, key comparisons, and meta-analysis,” *Metrologia* **54** S34-S62, 2017.

[22] A Possolo, A Koepke, D Newton, M R Winchester, “Decision Tree for Key Comparisons,” *Journal of Research of the National Institute of Standards and Technology* **126**:126007, 2021.

[23] J Mandel, “The validation of measurement through interlaboratory studies,” *Chemometrics and Intelligent Laboratory Systems* **11** 109-119, 1991.

[24] A Possolo, "Tracking truth through measurement and the spyglass of statistics," *Statistical Science* **38** (4): 655-671, 2023.

[25] O Bodnar, C Elster, "On the adjustment of inconsistent data using the Birge ratio," *Metrologia* **51** 516, 2014.

[26] S S Shapiro, M B Wilk "An analysis of variance test for normality," *Biometrika* **52** (3,4): 591–611, 1965.

[27] W Miao, Y R Gel, J L Gastwirth, "A new test of symmetry about an unknown median," *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, eds A C Hsiung, Z Ying, C H Zhang, World Scientific Singapore, pp 199–214, 2006.

[28] JCGM 100:2008, "Evaluation of measurement data – guide to the expression of uncertainty in measurement," (BIPM) 2008.

[29] M Borenstein, L V Dedges, J P T Higgins, H R Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," *Research Synthesis Methods* **1** 97–111, 2010.

[30] B Toman, A Possolo, "Laboratory effects models for interlaboratory comparisons," *Accreditation and Quality Assurance*, **14** 10 553–563, 2009.

[31] J Hartung, G Knapp, B K. Sinha, "Statistical Meta-Analysis with Applications", John Wiley & Son, Hoboken, NJ ISBN 978-0-470-29089-7, 2008.

[32] A Possolo, O Bodnar, "Redefining standard measurement uncertainty," *Ukrainian Metrological Journal* **1** 34–37, 2022.

[33] T P Hettmansperger, S J Sheather, "Confidence intervals based on interpolated order statistics," *Statistics & Probability Letters*, **4**(2) 75–79, 1986.

[34] R DerSimonian, N Laird "Meta-analysis in clinical trials," *Controlled Clinical Trials,* **7**(3) 177–188, 1986.

[35] R C Paule, J Mandel, "Consensus values and weighting factors," *Journal of Research of the National Bureau of Standards* **87** 377–385, 1982.

[36] D R White, "On the analysis of measurement comparisons," *Metrologia* **41** 122–131, 2004.

[37] S R Searle, G Casella, C E McCulloch, "Variance components," John Wiley & Sons, New York, ISBN 978-0-470-00959-8, 1992.

[38] W Vlechtbauer, "Conducting meta-analyses in R with the metafor package," *Journal of Statistical Software* **36**(3):1:48, 2010.

[39] A Rukhin, B Biggerstaff, M Vanghel, "Restricted maximum likelihood estimation of a common mean and the Mandel-Paul algorithm," *Journal of Statistical Planning and Inference* **83** 319–330, 2000.

[40] A Maronna, R Douglas Martin, V J Yohai, "Robust Statistics: Theory and Methods" (2nd edition), John Wiley & Sons, New York, ISBN 978-1-119-21466-3, 2018.

Appendix I Overview of the Analysis of the Past CCT Key Comparisons

CCT-K1 [7]

In this KC, 7 participants transferred their ITS-90 realizations in the temperature range from 0.65 K to 24.5561 K (^{3}He/^{4}He vapour pressure thermometers and ^{3}He/^{4}He interpolating constant-volume gas thermometers) to Rh-Fe resistance thermometers, delivered one or two calibrated Rh-Fe thermometers to the pilot, and all the Rh-Fe thermometers were intercompared in the pilot's cryostat at 43 different temperatures.

For each Rh-Fe thermometer, the participants provided the interpolation equation $T_i = T_i(R_i)$ and the calibration standard uncertainty u_i. The experimental uncertainty on the measurement made in the comparison cryostat, estimated by the pilot, was u_{exp}. The combined standard uncertainty of participant i for the comparison was then $u_{Ci} = \sqrt{u_i^2 + u_{\text{exp}}^2}$.

The bilateral equivalence between participants i and j was simply $D_{ij} = T_i - T_j$ with standard uncertainty $u(D_{ij})$ given by $u(D_{ij}) = \sqrt{u_i^2 + u_j^2 + 2u_{\text{exp}}^2}$, which assumes that there is no correlation between u_{Ci} and u_{Cj}.

The KCRV was defined as the weighted mean with weights $u_{Ci} = \sqrt{u_i^2 + u_{\text{exp}}^2}$:

$$T_{\text{KCRV}} = \frac{\sum_i \left(T_i / \sqrt{u_i^2 + u_{\text{exp}}^2} \right)}{\sum_i \left(1 / \sqrt{u_i^2 + u_{\text{exp}}^2} \right)}$$

The participants agreed that no additional significance was attached to the KCRV with respect to 'best estimates' of T_{90}. KCRV uncertainty was not computed.

CCT-K2 [8]

In this KC, each of the 7 participants calibrated 2 capsule-type SPRTs (CSPRTs) at the ITS-90 fixed points in the temperature range 13.8 K to 273.16 K, delivered the calibrated CSPRTs to the pilot, and all the CSPRTs were intercompared in the pilot's cryostat at temperatures close to the 8 low-temperature defining fixed points of the ITS-90.

For each CSPRT, the participants provided the resistance ratios $W(T_i)$ measured at the 8 calibration points, and the corresponding calibration standard uncertainty u_i. The experimental uncertainty on the measurement made in the comparison cryostat, estimated by the pilot, was u_{exp}. The combined standard uncertainty of participant i for the comparison was then $u_{Ci} = \sqrt{u_i^2 + u_{\text{exp}}^2}$.

The bilateral equivalence between participant i and j was simply $D_{ij} = T_i - T_j$ with standard uncertainty $u(D_{ij})$ given by $u(D_{ij}) = \sqrt{u_i^2 + u_j^2 + 2u_{\text{exp}}^2}$, which assumes that there is no correlation between u_{Ci} and u_{Cj}.

The KCRV was defined as the weighted mean with weights $u_{Ci} = \sqrt{u_i^2 + u_{\text{exp}}^2}$:

$$T_{\text{KCRV}} = \frac{\sum_i \left(T_i / \sqrt{u_i^2 + u_{\text{exp}}^2} \right)}{\sum_i \left(1 / \sqrt{u_i^2 + u_{\text{exp}}^2} \right)}$$

The KCRV represents "a common baseline against which all laboratory values can be compared". KCRV uncertainty was not computed.

CCT-K3 [9]

In this KC, 16 participants calibrated long-stem SPRTs (LSPRTs), provided by the pilot, at the ITS-90 defining fixed points in the range 83.8058 K to 933.473 K, and all SPRTs were measured at the pilot's fixed points.

For each LSPRT, the participants provided the resistance ratios $W(T_i)$ measured at the 7 fixed points, and the corresponding calibration combined standard uncertainty u_i. The participants were asked to report the Type A uncertainty ($S_{\text{A}i}/\sqrt{n_i}$ with $S_{\text{A}i}$ the repeatability for n_i realizations), the Type B uncertainty $S_{\text{B}i}$, and the corresponding degrees of freedom $\nu_{\text{A}i}$ and $\nu_{\text{B}i}$. The combined standard uncertainty u_i was thus given by: $u_i = \sqrt{S_{\text{A}i}^2/n_i + S_{\text{B}i}^2}$ and its degrees of freedom are obtained by applying the Welch–Satterthwaite formula (see [28]).

The bilateral equivalence between two participants i and j that measured only one and the same LSPRT was given by:

$$D_{ij} = T_i - T_j = \frac{W_i - W_j}{dW_{\text{r}}/dT}$$

$$U(D_{ij}) = \sqrt{k_i^2 (S_{\text{A}i}^2/n_i + S_{\text{B}i}^2) + k_j^2 (S_{\text{A}j}^2/n_j + S_{\text{B}j}^2) + k_{\text{SPRT}}^2 u_{\text{SPRT}}^2}$$

where u_{SPRT} is the uncertainty due to the drift of the LSPRT, and $k_{i,j,\text{SPRT}}$ are the coverage factors obtained from the t-distribution with $\nu_{i,j,\text{SPRT}}$ degrees of freedom: $k_i = t_{0.975}(\nu_i)$.

When participants i and j measured different LSPRTs, the bilateral equivalence was given by:

$$D_{ij} = (T_i - T_{\text{Pilot}}) - (T_j - T_{\text{Pilot}})$$

$$U^2(D_{ij}) = k_i^2(S_{\mathrm{A}i}^2/n_i + S_{\mathrm{B}i}^2) + k_j^2(S_{\mathrm{A}j}^2/n_j + S_{\mathrm{B}j}^2) + k_{\mathrm{NIST}}^2(S_{\mathrm{ANIST}}^2/n_{\mathrm{NIST1}} + S_{\mathrm{ANIST}}^2/n_{\mathrm{NIST2}}) + k_{\mathrm{SPRT1}}^2 u_{\mathrm{SPRT1}}^2 + k_{\mathrm{SPRT2}}^2 u_{\mathrm{SPRT2}}^2$$

It was decided that KCRV would not be given for this report. Although a minority of participants (5) thought that KCRV should be used, they withdrew their objections in order for the report to be completed.

CCT-K4 [10]

In this KC, the pilot circulated Al and Ag cells to the 3 co-pilots, and the co-pilots circulated their own Al and Ag cells to their loop participants (12 participants in total). The analysis approach was similar to that of CCT-K3. The expanded uncertainties were calculated using a coverage factor $k_i = t_{0.95}(\nu_i)$ from the t-distribution and the number of degrees of freedom for Type A uncertainties. Type B uncertainties were assumed to have an infinite number of degrees of freedom.

Given the comparison scheme, three different cases were considered. In the simplest case, in which the cell measured by participant *i* is the reference cell of participant *j*:

$$U(T_i - T_j) = \sqrt{U_i^2 + U_j^2 - t_{0.95}^2(\nu_j)u_{\mathrm{A}j}^2 + U_C^2}$$

where U_C is the uncertainty associated to the transport of the cell.

In the case in which participant *i* and participant *j* belongs to the same loop but none of the two is co-pilot, the expanded uncertainty is:

$$U(T_i - T_j) = \sqrt{U^2(T_i - T_{\mathrm{cop}}) + U^2(T_j - T_{\mathrm{cop}}) - 2U_{\mathrm{Bcop}}^2}$$

where cop in the subscripts refers to co-pilot and $U^2(T_{i,j} - T_{\mathrm{cop}})$ are calculated using the previous equations.

In the most general case, in which participant *i* and participant *j* belong to two different loops and none of the two is co-pilot, the expanded uncertainty is:

$$U^2(T_i - T_j) = U^2(T_i - T_{\mathrm{cop}i}) + U^2(T_{\mathrm{pil}} - T_{\mathrm{cop}i}) + U^2(T_j - T_{\mathrm{cop}j}) + U^2(T_{\mathrm{pil}} - T_{\mathrm{cop}j}) - 2U_{\mathrm{Bcop}i}^2 - 2U_{\mathrm{Bcop}j}^2 - 2U_{\mathrm{Bpil}}^2$$

where pil in the subscripts refers to pilot.

The KCRV was calculated as the weighted mean of all the differences $T_i - T_{\mathrm{pil}}$, with weights $u_i^2 + u_C^2$:

$$T_{\mathrm{KCRV}} = \frac{\sum_i \left(\left(T_i - T_{\mathrm{pil}} \right) / \left(u_i^2 + u_C^2 \right) \right)}{\sum_i \left(1 / \left(u_i^2 + u_C^2 \right) \right)}$$

The degrees of equivalence were defined by:

$$D_i = T_i - T_{\mathrm{KCRV}}$$

$$u(D_i) = \sqrt{u_i^2 + u_C^2}$$

CCT-K5 [11]

Two sets of two vacuum tungsten-strip lamps were circulated among the 14 participants in two geographical loops separated by pilot and assistant pilot measurements. The participants copied their ITS-90 local scales to the transfer lamps at 11 temperatures in the range 960 °C to 1700 °C. The results, obtained in the two separate loops on two different set of lamps were linked through the pilot and assistant pilot measurements, who measured both sets of lamps.

For the calculation of the KCRV, some measurements were not included in the KCRV calculation due to problems with the measurement system. The median was chosen for calculating the KCRV.

CCT-K6 [12]

A pair of commercial chilled-mirror hygrometers (CMHs) was circulated among the 10 participants in two geographical loops separated by pilot measurements. The participants generated 5 dewpoint temperatures $t_{\mathrm{ref}i}$ from –50 °C to 20 °C using their primary dewpoint generators and measured them using the travelling CMHs. The dew-point value t_i indicated by the travelling CMHs was the mirror PRT resistance, converted to a nominal temperature using IEC 60751 resistance-temperature characteristic. Each participant reported $T_i = t_{\mathrm{ref}i} - t_i$ and the corresponding combined standard uncertainty u_i.

The bilateral degrees of equivalence were given by:

$$D_{ij} = T_i - T_j = \left(t_{\mathrm{ref}i} - t_i \right) - \left(t_{\mathrm{ref}j} - t_j \right)$$

$$u\left(D_{ij} \right) = \sqrt{u_i^2 + u_j^2 + u_{\mathrm{drift}}^2}$$

$$U\left(D_{ij} \right) = 2u\left(D_{ij} \right)$$

The KCRV was calculated as the weighted mean of all results:

$$T_{\mathrm{KCRV}} = \frac{\sum_i \left(\left(T_i \right) / \left(u_i^2 \right) \right)}{\sum_i \left(1 / u_i^2 \right)}$$

$$\frac{1}{u^2(T_{\mathrm{KCRV}})} = \sum_i \frac{1}{u_i^2}$$

The results of one of the participants at two temperatures were excluded from the calculation of the KCRV. It was noted that the KCRV has no absolute significance outside the comparison, though it is essential to the use of comparison results for review of CMC claims.

CCT-K7 [13]

In this KC, each of the 21 participants calibrated a transfer cell against their national reference for the triple point of water and delivered it to the pilot. The pilot intercompared all transfer cells.

Indicating by $T_{\mathrm{tr}i}$ the temperature realized by the transfer cell of participant *i* and by $T_{\mathrm{ref}i}$ the temperature realized by the national reference of participant *i*, each participant measured $T_{\mathrm{tr}i} - T_{\mathrm{ref}i}$ with standard uncertainty u_i.
The pilot measured the differences $T_{\mathrm{tr}i} - T_{\mathrm{refpilot}}$. From the participants' measurements and the pilot's measurements, the differences $T_{\mathrm{ref}i} - T_{\mathrm{refpilot}}$ could be calculated, with uncertainties:

$$u^2\left(T_{\mathrm{ref}i} - T_{\mathrm{refpilot}}\right) = u^2\left(T_{\mathrm{tr}i} - T_{\mathrm{refpilot}}\right) + u^2\left(T_{\mathrm{tr}i} - T_{\mathrm{ref}i}\right)$$

The uncertainties in the two differences were considered totally uncorrelated.

The KCRV was calculated as the simple mean of the results from all the participants:

$$T_{\mathrm{KCRV}} - T_{\mathrm{refpilot}} = \frac{1}{N}\sum_i \left(T_{\mathrm{ref}i} - T_{\mathrm{refpilot}}\right)$$

The uncertainty of the KCRV was calculated as the standard deviation of the mean:

$$u\left(T_{\mathrm{KCRV}} - T_{\mathrm{refpilot}}\right) = \sqrt{\frac{\sum_i u^2\left(T_{\mathrm{ref}i} - T_{\mathrm{refpilot}}\right)}{N-1}}$$

CCT-K8 [14]

A pair of commercial chilled-mirror hygrometers (CMHs) was circulated among the 10 participants in two loops. The participants generated 7 dewpoint temperatures $t_{\mathrm{ref}i}$ from 30 °C to 95 °C using their primary dewpoint generators and measured them using the travelling CMHs. The dew-point value t_i indicated by the travelling CMHs was the mirror PRT resistance, converted to a nominal temperature using IEC 60751 resistance-temperature characteristic. Each participant reported $T_i = t_{\mathrm{ref}i} - t_i$ and the corresponding combined standard uncertainty u_i.

The bilateral equivalence between participant *i* belonging to loop loop*i* and participant *j* belonging to loop loop*j* was obtained as:

$$D_{ij} = (t_i - t_{\mathrm{ref}i})_{\mathrm{loop}i} - \left(t_j - t_{\mathrm{ref}j}\right)_{\mathrm{loop}j} + B_{ij}$$

The term B_{ij} is a discrete function of *i* and *j*, used to link the two loops, obtained from the measurements at the linking labs.

Because the two loops measured different travelling standards, it was necessary to generate a virtual transfer standard having an intermediate behaviour between the two real travelling standards. This was done by applying a correction factor $C_{\mathrm{loop}i}$ to both loops (the fact that the linking labs measured all travelling artifacts 4 times made it possible to calculate the correction factors):

$$C_{\mathrm{loop1}} = B_{ij}/2$$

$$C_{\mathrm{loop2}} = -B_{ij}/2$$

$$u(C_{\mathrm{loop}i}) = \frac{1}{2}u(B_{ij})$$

The KCRV was calculated as the weighted mean of all results with one exception: at the 3 highest temperatures, the results of one participant were excluded from the calculation of the KCRV, because its standard was metrologically traceable to another participant. However, the KCRV has no absolute significance.

CCT-K9 [15]

Fourteen NMI laboratories participated in this comparison. Each participating NMI selected two of their own SPRTs, calibrated them at every available ITS-90 fixed point within the range 83.8058 K to 692.677 K, and transported them to the pilot laboratory (NIST). Following the NIST calibration, the SPRTs were returned to their respective owners, and each of them recalibrated the SPRTs at their facility.

The calibrations were performed in terms of resistance ratio *W* at each fixed point. For each fixed point, each participant *i* measured the initial values of *W* for their two SPRTs: W_{i1}^{Ante} and W_{i2}^{Ante}. The SPRTs were then transported to and measured at NIST: W_{i1}^{Pilot} and W_{i2}^{Pilot}. Finally, the SPRTs were returned to their respective NMIs for a final measurement: W_{i1}^{Post} and W_{i2}^{Post}.

The result of participant *i* was then (*j* identifies either SPRT 1 or SPRT 2):

$$\bar{W}_{ij} = \frac{1}{2}\left(W_{ij}^{\mathrm{Ante}} + W_{ij}^{\mathrm{Post}}\right)$$

The temperature difference ΔT_{ij} between participant *i* and the pilot, calculated for SPRT *j* was:

$$\Delta T_{ij} = \frac{\bar{W}_{ij} - \bar{W}_{ij}^{\mathrm{Pilot}}}{dW_{\mathrm{r}}/dT} + C_{ij}$$

where C_{ij} accounts for the drift of the SPRT.

At each fixed point and for each participant *i*, the ΔT_{ij} are combined to give ΔT_i:

$$\Delta T_i = \frac{1}{2}(\Delta T_{i1} + \Delta T_{i2})$$

The KCRV $\overline{\Delta T}$ was calculated as:

$$\overline{\Delta T} = \frac{1}{n}\sum_{i=1}^{n} \Delta T_i$$

The SPRTs that failed the cut off criteria were excluded from the calculation of the KCRV.

The DoEs were calculated as:

$$D_i = \Delta T_i - \overline{\Delta T}$$

The uncertainty in the KCRV and in the DoEs were calculated in the same way of CCT-K3. Systematic uncertainties were treated as completely correlated and combined according to GUM section 5.2.2, the most conservative approach.

CCT-K10 [16]

CCT-K10 consisted of two parts:

1. Two transfer radiation thermometers and a transportable copper fixed-point blackbody source were circulated among the participants to compare their ITS-90 scale realizations at 14 temperatures over the range from the Ag point to 3000 °C.
2. Three high temperature fixed-point (HTFP) blackbody cells were circulated among the participants to probe their scale realizations at these particular temperatures.

The transfer instruments were circulated in the form of a semi-collapsed star (or flower), in which the artefacts were sent from the pilot laboratory to one RMO for a loop around that region, and then the artefacts were sent back to the pilot laboratory. Following check measurements at the pilot laboratory, the artefacts were then sent to the next RMO region for a further loop.

Initially the analysis of the results was carried out according to the method described in the protocol (weighted mean with cut off). However, due to the presence of some outlying data, and after further discussion amongst the participants, it was decided to use the median as the reference value (including the average pilot results) for the part of the comparison using radiation thermometers. For the comparison with HTFPs, the reference value was calculated as the weighted mean with cut off, as specified in the protocol.

The uncertainty *u* in the median value was calculated using the 'median of absolute deviations' (or 'MAD') as:

$$u \cong \frac{1.858}{\sqrt{(n-1)}}\text{MAD}$$

where n is the number of results and

$$\mathrm{MAD} = \mathrm{median}\{|x_i - \mathrm{med}|\}$$

where x_i is the result of laboratory i and med is the median value.

The DoEs were calculated for each participant for each thermometer and for each nominal temperature, using the difference between the result of each participant and the reference value (median) for that particular thermometer, and the uncertainty of the difference, which is the participant uncertainty (including a component for the thermometer drift) combined with the uncertainty of the reference value. The pairwise equivalence between each pair of participants was also determined.