METRICA
METRology Institutional CAtalog

ISTITUTO NAZIONALE DI RICERCA METROLOGICA
Repository Istituzionale

Conformance probability in the assessment of Calibration and Measurement Capabilities

Original
Conformance probability in the assessment of Calibration and Measurement Capabilities / Malengo, A., Bich, W.. - In: MEASUREMENT. - ISSN 0263-2241. - 192:(2022), p. 110865. [10.1016/j.measurement.2022.110865]

Availability:
This version is available at: 11696/76165 since: 2023-02-28T13:31:35Z

Publisher:
ELSEVIER SCI LTD

Published
DOI:10.1016/j.measurement.2022.110865

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

14 October 2026

Measurement 192 (2022) 110865

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Conformance probability in the assessment of Calibration and Measurement Capabilities

Andrea Malengo [*,1], Walter Bich [1]

Istituto Nazionale di Ricerca Metrologica, INRiM, Strada delle Cacce 91, 10135 Torino, Italy

ARTICLE INFO

Keywords:
Interlaboratory comparisons
Calibration and measurement capabilities
Conformity assessment

ABSTRACT

We argue that the assessment of the Calibration and Measurement Capabilities, CMCs, by means of the results of a Key Comparison is a *bona fide* exercise of conformity assessment, and as such should be treated, using the appropriate tools, including risk assessment. This position contrasts with the current practice, in which acceptance or rejection of a CMC claim are based on the normalised error. We show that, behind this seemingly unique acceptance criterion, different decision rules – guarded acceptance, simple acceptance and guarded rejection – exist in reality, depending on the characteristics of the comparison. This variety of decision rules impairs the fairness of the current equivalence arrangement. We suggest that the conformance probability should be the key parameter to be considered in the assessment of a CMC claim. Using a suitable Probability Density Function, PDF, for the measurand, we calculate the conformance probability for the possible scenarios, and show that using the current acceptance criterion the conformance probability can attain unacceptably low values. Therefore, we maintain that the current acceptance criterion is ambiguous and inadequate, and suggest to rather adopt a criterion based on the calculation of the conformance probability and the establishment of a minimum threshold for acceptance. We demonstrate our proposal by applying it to a practical case and to a fictitious example in mass metrology.

1. Introduction

The Mutual Recognition Arrangement of the International Committee for Weights and Measures (CIPM MRA) [1] establishes the interlaboratory comparisons (CIPM or regional) as the master tool to demonstrate evidence of the Calibration and Measurement Capabilities (CMCs) declared by National Metrology Institutes (NMIs) for their uploading on the BIPM key comparison database, KCDB [2], their authoritative repository. Also, comparisons are carried out at suitable time intervals to check that the performance is maintained. Similarly, within the framework of the ILAC Mutual Recognition Arrangement (ILAC MRA) [3], interlaboratory comparisons are regularly carried out among accredited calibration laboratories and a reference laboratory (typically but not necessarily the local NMI) with similar purposes. In this paper, we focus on CIPM key comparisons (KCs) for simplicity and with no loss of generality. Indeed, most of the following considerations apply as well to any generic interlaboratory comparison (ILC).

Guidance exists [4] on how to draw a protocol for and how to perform a KC. There is as well a huge literature concerning more or less sophisticated statistical methods to obtain from a data set a consensus, or reference, value (KCRV) and the associated uncertainty (see some references in Section 3 below). In some fortunate cases (for example in chemistry and ionising radiation), the KCRV may be given by a primary method, thus representing a reference external to the data set, with an associated uncertainty often negligible compared to those of the participants. A similar situation holds in the field of proficiency tests, where international standards provide specific guidance [5,6].

Whatever the case, the degrees of equivalence (DoEs), unilateral or bilateral, are thus obtained as the difference between the estimate of the participating laboratory and the KCRV or another participant, respectively, and the expanded uncertainty of the difference. Hereafter, we will only consider unilateral DoEs.

To the best of our knowledge, there exists no clear prescription on how to infer or confirm the CMCs based on the DoEs and more generally on the outcome of a KC. Also the relevant literature is meagre (see Section 2 for some instances). All that exists is a more or less consolidated practice, universally adopted, perhaps with some variation, in international and regional KCs, in supplementary comparisons and in the ILCs carried out in the framework of accreditation.

* Corresponding author.
E-mail addresses: a.malengo@inrim.it (A. Malengo), w.bich@inrim.it (W. Bich).
[1] Both authors contributed equally to the manuscript.

https://doi.org/10.1016/j.measurement.2022.110865
Received 16 December 2021; Received in revised form 23 January 2022; Accepted 6 February 2022
Available online 17 February 2022
0263-2241/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

We argue that the process of inferring or confirming a CMC based on the result of a comparison is a *bona fide* conformity assessment, and as such should be treated.

We also argue that the DoE as a tool to validate a claimed CMC is inappropriate, as the two concepts do not have a clear connection.

In this paper, we first recall DoEs and expound our understanding of a CMC (Section 2). We then discuss the concept of consistency (Section 3), and illustrate the practice currently used to validate a CMC claim (Section 4). We present our proposal in Section 5 and apply it to a real KC in Section 6. We also discuss the implications for the future mass comparisons (Section 7). Section 8 concludes the paper.

2. Degree of equivalence and calibration and measurement capability

2.1. Degree of equivalence

The CIPM Mutual Recognition Arrangement, CIPM MRA, [1] was signed in 1999. A set of ancillary guidance documents superseding previous documents has been published in 2021 [4,7]. The part of the MRA relevant here is the technical supplement (revised in 2003), where in section T.2 the *degree of equivalence* (DoE henceforth) is defined as

> The degree of equivalence of each national measurement standard is expressed quantitatively by two terms: its deviation from the key comparison reference value and the uncertainty of this deviation (at a 95 % level of confidence).

This definition is perfectly clear, but no hints are given on what to do with DoEs. As a result, no agreed procedure exists on how to use DoEs to support a CMC claim, especially when a laboratory is not consistent with the KCRV. To the best of our knowledge, the only documented attempt to assign a CMC in a technically sustainable way when a participant is not consistent with the KCRV is given in [8], where a procedure is suggested (and possibly adopted by some Consultative Committees, such as the CCQM) that is not too different from the practice commonly adopted by most CCs (see Section 4 below). A proposal dealing with the global expansion of the CMC uncertainties over a whole range of measurands [9] is beyond the scope of this paper.

2.2. Calibration and measurement capability

The definition of Calibration and Measurement Capability, CMC, elaborated in a paper by a BIPM/ILAC joint working group [10] and also given in other documents [7,11], reads:

> a CMC is a calibration and measurement capability available to customers under normal conditions...

and further on (in [7]):

> In the KCDB, a CMC is characterised by the measured quantity and associated expanded measurement uncertainty (generally given at a 95 % coverage level of confidence), for a given range...

In practice, the declared uncertainty is the best measurement uncertainty that can be expected by the laboratory for a given measurand.

Suppose a laboratory has in the KCDB a registered CMC for a nominal quantity value X with an expanded uncertainty U_x. This declaration means that the laboratory is internationally recognised as being able to produce consistently estimates x_j of that nominal quantity value X such that there is a 95 % chance that the true values X_j lie within $x_j - U_x$ and $x_j + U_x$.

The same considerations apply to Best Measurement Capabilities, BMCs, a term used historically in connection with the uncertainties stated in the scope of an accredited laboratory, the term having the same meaning as CMC [11].

3. Consistency – A reminder

Any estimate y (*measured value* in the VIM [12]) of a measurand Y leaves a *penumbra of uncertainty* ([13], p. 33) about the true value of Y. The penumbra is typically expressed by a standard, or expanded, measurement uncertainty $u(y)$ or $U(y)$, respectively. The concepts of estimate, measurand, standard and expanded uncertainty are defined and universally understood. The same is not true for *consistency*, yet, a key concept in modern worldwide metrology and the driving force behind any international comparison. Wherever two or more measurement results (*i.e.*, estimates and associated uncertainties) of the same measurand are involved, the key question is whether they are *consistent* or not. There is not a formal definition of consistency, although the term is generally understood as a synonym for *compatibility*, this term being formally defined (see [12], definition 2.47). Even worse, consistency has a main technical meaning in statistics (see, *e.g.*, [14–16]) which differs from what a metrologist typically intends for it.

Using matrix notation, we consider a set of measured values $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^\top$, viewed as realisations of random variables $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^\top$, to be consistent if the random variables $\boldsymbol{X}$ have the same expectation μ. In other words, a consistent data set obeys the statistical model

$$\boldsymbol{x} = \mu\mathbf{1} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{1} = (1, 1, \dots, 1)^\top$ and $\boldsymbol{\epsilon}$ is a vector of unknown deviations, taken as realisations of random variables $\boldsymbol{E}$ having expectation equal to zero, *i.e.*, $\mathrm{E}(\boldsymbol{E}) = \mathbf{0}$ and covariance matrix

$$\boldsymbol{U}(\boldsymbol{\epsilon}) = \boldsymbol{U}(\boldsymbol{x}). \tag{2}$$

The further, realistic assumption is usually made that the X_i are normally distributed $X_i \sim \mathrm{N}(\mu, u_i^2)$.

The adequacy of model (1) for a specific data set must be checked against some criterion. There exists a variety of different criteria, either optimal or robust [17,18] which, when applied to a specific data set, can produce different responses. The criterion by far more adopted in metrology is the well-known chi-squared, or χ^2 test [19,20].

In a general form, the test is considered passed if

$$\Pr\{\chi^2(\nu) > \chi^2_{\mathrm{obs}}\} > \alpha, \tag{3}$$

where Pr denotes probability, $\nu = n - 1$ is the degrees of freedom, the observed statistic χ^2_{obs} is

$$\chi^2_{\mathrm{obs}} = \left(\boldsymbol{x} - \hat{\mu}\mathbf{1}\right)^\top \boldsymbol{U}(\boldsymbol{x})^{-1} \left(\boldsymbol{x} - \hat{\mu}\mathbf{1}\right), \tag{4}$$

$\hat{\mu}$ being an estimate of μ, and α is a value suitably chosen. The left-hand side of inequality (3) is the so-called p-value, the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It can be broadly related, with many *caveat* [21–25], to the probability of being wrong in rejecting the null hypothesis, in this case the hypothesis of randomness of the data set, expressed by model (1).

In terms of quantiles, expression (3) is written as

$$\chi^2_{\mathrm{obs}} \le q_{1-\alpha}, \tag{5}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2(\nu)$ distribution.

In the context of the adjustment of fundamental constants [26], $\chi^2_{\mathrm{obs}} \le \nu + \sqrt{2\nu}$. In that of intercomparisons, $\alpha = 0.05$ [20], so that condition (5) is $\chi^2_{\mathrm{obs}} \le q_{0.95}$.

In the common case in which an external estimate of the true value of the measurand is not available, the typical choice for the estimate $\hat{\mu}$ (commonly denoted as x_{ref} in the context of KCs) of the measurand is the weighted mean, which is optimal among linear estimators if data agree with model (1), independently of their probability distribution. The weighted mean is the optimal estimator if data are normally distributed [15].

The imagination of physicists, metrologists and statisticians has produced a great number of different estimators for the common case in which data do not behave according to model (1). This reflects the widespread interest of the topic in the most disparate fields, from the adjustment of fundamental constants [27–30] to the comparisons of standards in metrology [31–34] to meta-analysis in medicine [33,35–38]. A general tool is provided in [39,40].

4. CMC assessment – Current practice

The validation of a CMC is a complex process involving, among others, considerations about the quality management system of the candidate laboratory. In any case, the fundamental step to demonstrate evidence of a claimed CMC is to evaluate the performance of the laboratory in a KC. In this paper, we restrict the discussion to this aspect. In the current practice, the performance is in most cases evaluated by using the two parts of which the DoE is made so as to form the so-called *normalised error* E_n [5,6,41]:

$$E_n = \frac{x - x_{ref}}{U\left(x - x_{ref}\right)} = \frac{\Delta x}{U_{\Delta x}}, \tag{6}$$

where Δx (we dropped the i index with no loss of generality) is the deviation of the laboratory result x from the Key Comparison Reference Value (KCRV) x_{ref} and $U_{\Delta x}$ is the associated expanded uncertainty at the 95 % coverage probability. Assuming normality, the denominator is in general

$$U_{\Delta x} = 2\sqrt{u_x^2 + u^2(x_{ref}) - 2u(x, x_{ref})}, \tag{7}$$

the coverage factor $k = 2$ rather than $k = 1.96$ being used, according to the customary practice.

The covariance $u(x, x_{ref})$ (or $\rho u_x u(x_{ref})$, where ρ is the linear correlation coefficient) plays a key role. Whilst its presence in the denominator is obvious nowadays, it was not so in the past. It was indeed thanks to a remark by W Woeger [42] that the covariance was included in calculation. In that paper, references are given demonstrating that the customary practice at that time was to calculate E_n without covariance.

When $U_x = 2u_x$ is equal to the CMC claim of the laboratory, as is typically the case, and as we will assume henceforth, the claim is considered correct if $|E_n| \le 1$.

When a laboratory is not consistent, i.e., when the normalised error $|E_n| > 1$, the common practice is simply to enlarge the uncertainty U_x of the CMC so that it is (at least) equal to Δx, or to adopt the procedure suggested in [8].

The statistical significance of the requirement $|E_n| \le 1$ can be understood considering that the normalised error is related to the chi-squared statistic for $n = 2$ by $4E_n^2 = \chi^2_{obs}$ [see Appendix, Eq. (14)].

Therefore, the requirement $|E_n| \le 1$ translates into the requirement $\chi^2_{obs} \le 4$. Or, $q_{1-\alpha} = 4$ for the $\chi^2(1)$ distributions corresponds to $1 - \alpha = 0.954$. This means that, under the assumption that x and x_{ref} estimate the same measurand, *i.e*, that they are consistent, the probability of finding a difference equal to or greater than Δx is a mere 4.6 %.[2] Of course, this does not mean that there is a 95.4 % probability that they are consistent, let alone that the laboratory conforms to the claim. Nonetheless, if the $|E_n| \le 1$ test is passed, the CMC claim of the laboratory is considered correct. This means, as recalled in Section 2, that the laboratory, provided that all the further necessary conditions are met, will be internationally recognised as being capable of measuring measurands similar to X in such a way that its estimates are 'close' to the true values of those measurands.[3] Or, to say better, that there is a 95 % probability that the true values of the measurands lie within the interval spanned by the estimate plus or minus the expanded uncertainty. In the following sections we will try to show that this weak acceptance criterion can lead to perverse consequences.

[2] To be compared with the more stringent requirement $\chi^2_{obs} \le \nu + \sqrt{2\nu}$ mentioned in Section 3, corresponding to a more reassuring value $1-\alpha = 0.880$.

[3] What exactly is meant by 'similar' is the topic of a never-ending debate (how far does the light shine etc.), fortunately beyond the scope of this paper.

5. CMC assessment – Our proposal

We argue that the validation of a claimed CMC is a true conformity assessment, and that the success in the test $|E_n| \le 1$, being just a (considerably weak) consistency test, is not adequate to validate meaningfully the CMC. Conformity assessment requires decision rules concerning the acceptance (or rejection), and involves risk evaluation. In this section we consider the CMC assessment from the viewpoint of a conformity assessment.

5.1. Conformity assessment – a reminder

To assess the conformity of an item with a specified requirement, according to the document that we consider as the gold standard in the field [43],

- the item is distinguished by a single scalar quantity;
- an interval $[T_L, T_U]$ of permissible values of the quantity is specified;
- the property can be measured and the measurement result … expressed in a manner consistent with the principles of the GUM…

Under these three conditions, the conformity assessment consists of the following steps:

- measure the property of interest;
- compare the measurement result with the specified requirement;
- decide on a subsequent action.

5.2. Key comparison as a conformity assessment

To position a KC exercise within the framework of conformity assessment, we assume

1. the item is the travelling standard, distinguished by the quantity it materialises;
2. the property of interest is, strictly speaking, the CMC claimed by that laboratory (typically U_x, the expanded uncertainty declared in the comparison exercise). In practice, the CMC is probed, so to say, by $\Delta X = x - X$, the deviation of the laboratory estimate from the true value of the measurand; it is estimated by $\Delta x = x - x_{ref}$.
3. the specified requirement is thus $-U_x \le \Delta X \le U_x$. The (two-sided) tolerance interval $[T_L, T_U]$ is $[-U_x, +U_x]$; it follows that $U_x = \left[T_U - T_L\right]/2 = T/2$, where T is the *tolerance* as defined in [43].
4. finally, we suggest the following (binary) decision rule: the claim of the laboratory is accepted if the *conformance probability* p_c (see Section 5.3) has at least a stipulated value; otherwise the claim is rejected and suitable corrective actions are to be taken.

The fourth point above is our proposed alternative to the current decision rule, *i.e.* that the claim is accepted if $|E_n| \le 1$.

5.3. Conformance probability

The conformance probability p_c is defined as the *probability that an item fulfils a specified requirement* ([43], definition 3.3.7). In the case of a CMC claim, it represents the probability that the deviation ΔX of the estimate provided by the laboratory from the true value is within the tolerance interval claimed by the laboratory itself. The conformance probability is thus the fraction of the state-of-knowledge probability distribution for $\Delta X = x - X$, where X is the true value, that falls within the tolerance interval $[-U_x, +U_x]$, *i.e.*

$$p_c = \int_{-U_x}^{+U_x} g_{\Delta X}(\xi)\mathrm{d}\xi = G_{\Delta X}\left(U_x\right) - G_{\Delta X}\left(-U_x\right), \tag{8}$$

where $g_{\Delta X}(\xi|x, x_{\mathrm{ref}})$ and $G_{\Delta X}(z) = \int_{-\infty}^{z} g_{\Delta X}(\xi)\mathrm{d}\xi$ are the probability density function (PDF) and the cumulative distribution function (CDF) for ΔX, respectively.

We adopt a Gaussian (normal) PDF for simplicity and for its widespread use. In Bayesian inference, and assuming a Gaussian likelihood, this PDF would be a reasonable approximation of the posterior for ΔX, given a non-informative prior (see, *e.g.* [43], 7.2). Other choices might be more appropriate in specific cases.

We thus write the PDF $g_{\Delta X}(\xi|x, x_{\mathrm{ref}})$ for ΔX as

$$g_{\Delta X}(\xi|x, x_{\mathrm{ref}}) = \mathbf{N}[\Delta x, u^2(x_{\mathrm{ref}})] = \frac{1}{\sqrt{2\pi}u(x_{\mathrm{ref}})} \exp\left[-\frac{1}{2}\left(\frac{\xi - \Delta x}{u(x_{\mathrm{ref}})}\right)^2\right]. \quad (9)$$

The PDF (9) has expectation Δx, the estimate of ΔX. As to the variance, we argue that, in the context of conformity assessment, the only uncertainty in ΔX is the one about X, for which only an estimate, x_{ref}, is available. Such a choice for the variance of the PDF (9) for ΔX epitomises the fundamental difference between a consistency test and a conformity assessment. In the former, in which two estimates (x_{ref} and x) of the same true value X are compared to check their consistency, the uncertainties associated with both estimates play similar roles. In the latter, the estimate x is checked against the true value to ascertain whether it fulfils the requirement, that is, whether its distance from the (uncertain) true value is smaller than the tolerance interval. In this second exercise, the (squared) uncertainty $u^2(x_{\mathrm{ref}})$ about the true value X constitutes the variance of the PDF for the measurand; the expanded uncertainty declared by the candidate laboratory, U_x, forms the tolerance interval. The two roles are distinct, and it would be meaningless to use u_x in both the tolerance interval, that is, the integration limits, and in the PDF for the measurand. It is worth recalling that an uncertainty is *associated with* the estimate, and is not the uncertainty *of* the estimate (as sometimes can be found in loosely written documents). There is no such a thing as the uncertainty about the estimate, there is rather, *associated* with the estimate, an uncertainty about the true value, and this latter uncertainty is irrelevant in this specific conformity assessment. This attitude is aligned with the general scheme of the Bayesian view of probability (see, *e.g.* [44,45]).

A more conventional, in our opinion mistaken choice, would be to include in the uncertainty the contribution u_x and $u(x, x_{\mathrm{ref}})$. In any case, the considerations of the next section hold independently of the choice of the PDF, the choice only affecting (mildly) the values of p_c.

5.4. Decision rules and conformance probability with the current E_n criterion

We discuss here the conformance probability (8) yielded by the current criterion, by which the CMC is validated if $|E_n| \leq 1$, *i.e.*, $|\Delta x| \leq U_{\Delta x}$, where $U_{\Delta x}$ is defined as in Eq. (7). In practice, the expanded uncertainty $U_{\Delta x}$ dictates the *acceptance interval* $[A_L, A_U] = [-U_{\Delta x}, U_{\Delta x}]$ for Δx (see [43], definition 3.3.9). $U_{\Delta x}$ can be greater or smaller than U_x, depending on whether the covariance is equal to zero or not, respectively (note that the covariance, and thus ρ, are always positive in comparisons). Accordingly, the acceptance interval can be greater or smaller than the tolerance interval $[-U_x, +U_x]$.

When $\rho = 0$ the acceptance interval is greater than the tolerance interval and the criterion $|E_n| \leq 1$ is equivalent to a *guarded rejection decision rule* (see [43], 8.3.3). This implies a negative *guard band* $w = U_x - U_{\Delta x}$ (see [43], 3.3.11), thus reducing the conformance probability and correspondingly increasing the *consumer's risk* $R_C^* = 1 - p_c$ (see [43], 9.3.2).

Vice versa, when $\rho > 0$ the acceptance interval is smaller than the tolerance interval, and the criterion $|E_n| \leq 1$ is equivalent to a *guarded acceptance decision rule* (see [43], 8.3.2). The guard band is here positive, thus increasing the conformance probability and reducing the consumer's risk.

In either cases, the amplitude of the guard band depends on $u(x_{\mathrm{ref}})$.

What is relevant here is that, with the current criterion, an apparently unambiguous decision rule, *i.e.*, that the CMC claim is validated if $|E_n| \leq 1$, corresponds in reality to different (and contrasting) decision rules, depending on elements other than the performance of the candidate laboratory (epitomised by Δx and u_x), such as ρ and $u(x_{\mathrm{ref}})$ (as concerns the severity of the decision rule). The unpleasant consequence is that the consumer's risk is non-homogeneous along the CMCs currently registered in the KCDB. Even worse, the risk is currently unknown, its calculation having been neglected so far.

A relevant case is when a random instability of the measurand causes an increase of $u(x_{\mathrm{ref}})$ and thus an increase of the guard band and of the consumer's risk. Some investigations [46,47] considered the situation described above in terms of loss of statistical power of the E_n criterion. Therefore, they remain in the context of consistency testing, and do not introduce conformity-assessment considerations. Our opinion in this respect is that the random instability should be viewed as an intrinsic uncertainty of the measurand. As such, it should affect all the estimates of the measurand, contrary to the current practice, in which it affects only the uncertainty of the KCRV. In our proposal, the additional uncertainty contributes both to the variance of the PDF for the measurand and to the tolerance interval. The resulting conformance probability seems to be less affected than the normalised error in this case, which deserves further investigation.

5.5. Worst-case conformance probabilities

We study here the worst-case conformance probability ($|E_n| = 1$) in various situations that occur in practice, assuming that the *measurement capability index* (see [43], 7.6.2) $C_m = u_x/u(x_{\mathrm{ref}}) \geq 1$, *i.e.*, that $u(x_{\mathrm{ref}})$ is never larger than u_x. This condition is a strict requirement, and holds true for all interlaboratory comparisons, be they KCs or bilateral comparisons between a reference laboratory and a laboratory candidate to accreditation.

Conformance probabilities, as given by the integral (8) with the PDF (9), were calculated using CaSoft [48,49], a dedicated software of intuitive usage. The results were double-checked using Wolfram Mathematica 13.0.

We consider separately the two cases, $\rho = 0$ and $\rho > 0$.

1. $\rho = 0$.
   This situation occurs when the estimate of the measurand is obtained from an external measurement (as frequently happens in gas-analysis KCs and almost invariably in proficiency tests, PTs).

   (a) When $u(x_{\mathrm{ref}}) \ll u_x$, the acceptance interval virtually coincides with the tolerance interval. In this scenario of *simple acceptance* (see [43], 8.2), the conformance probability runs from a minimum of 50 % for $|E_n| = 1$ (*i.e.*, $|\Delta x| \approx U_x$) to a maximum close to 100 % for $|E_n| = 0$ (*i.e.*, $|\Delta x| = 0$): $0.5 \leq p_c < 1$.
   We reasonably set a threshold $u(x_{\mathrm{ref}}) \leq u_x/3$ (*i.e.*, $C_m = 3$) for this case.[4] At this threshold, both $U_{\Delta x}$ and the acceptance interval increase by a negligible 5.4 %. The corresponding guard band is $w = -0.16U(x_{\mathrm{ref}})$, but the conformance probability, for $|E_n| = 1$ decreases to a little reassuring $p_c = 37$ %.

   (b) When $u(x_{\mathrm{ref}})$ is meaningful, say, $u(x_{\mathrm{ref}}) > 1/3u_x$, the decision rule is that of a guarded rejection, thus increasing the consumer's risk to the advantage of the laboratory. The guard bands are negative, and can be as great as $0.41U_x$ in the extreme case $u(x_{\mathrm{ref}}) = u_x$, *i.e.* $C_m = 1$.

[4] This is a common limit in legal metrology, see [43], EXAMPLE in 8.2.3.

A result yielding $|E_n| = 1$ would be accepted with a conformance probability of barely $p_c = 20\,\%$. This situation represents the limiting case for which a result of the comparison is accepted, and is the most favourable condition for the participating laboratory.
However, even for the more realistic value $u(x_{ref}) = 0.8u_x$ ($C_m = 1.25$), the conformance probability would still be a mere $p_c = 24\,\%$.

2. $\rho > 0$.
The effect of a finite (positive) correlation coefficient is to reduce $u_{\Delta x}$ and thus the acceptance interval. In this case, the criterion $|E_n| \leq 1$ is equivalent to a guarded acceptance, thus reducing the consumer's risk at the expenses of the laboratory.

   (a) The situation $u(x_{ref}) \ll u_x$ represents the ideal case of a well-behaving KC, in which the measurand is stable and its estimate is obtained as the weighted mean of a highly consistent data set. In this case, $\rho = u(x_{ref})/u_x$.
   At the threshold $C_m = 3$, both $U_{\Delta x}$ and the acceptance interval decrease by a negligible 5.4 % and the corresponding guard band is now positive: $w = 0.17U(x_{ref})$.
   As concerns conformance probability, for $|E_n| = 1$ it is $p_c = 63\,\%$.

   (b) When $u(x_{ref}) > u_x/3$, its impact on $U_{\Delta x}$ becomes significant.
   For example, considering the extreme yet realistic case $\rho = u(x_{ref})/u_x$ and $u_x \approx u(x_{ref})$, the uncertainty $U_{\Delta x} \approx 0$. This last situation represents the worst condition for the participating laboratory, yet it can occur i) in bilateral comparisons between laboratories with similar capabilities, where one laboratory not only provides the reference value but also the traceability to the other participating laboratory; or, ii) (less markedly) when in a KC the reference value is estimated by a few laboratories with similar uncertainty.
   As an example, if $u(x_{ref}) = 0.8u_x$ and $\rho = u(x_{ref})/u_x = 0.8$, the positive guard band would be $w = 0.5U(x_{ref})$, and the conformance probability $p_c = 84\,\%$.
   This situation may arise in bilateral as well as in multilateral comparisons with a small number of participants, or in comparisons for accredited laboratories.

As a general consideration, when $u(x_{ref})$ is meaningful, the guard bands, positive or negative depending on ρ, increase in width and the difference between tolerance and acceptance intervals can become so large that the consumer's risk attains beyond-reasonable (small or great) values.

Table 1 summarises the minimum possible conformance probabilities for the various cases considered.

As a last remark, consider a laboratory A obtaining traceability from a laboratory B and willing to validate its CMC by participating in a bilateral comparison. For A it is more convenient to perform the comparison with a third laboratory C (having uncertainties comparable to those of B) rather than with B itself. In the former case there is no correlation and the decision rule is a guarded rejection, thus favourable to the laboratory. In the latter, the decision rule is a guarded acceptance. The difference may be decisive as concerns the acceptance or rejection of the claim.

5.6. Decision rule

As already mentioned in Section 5.2, we propose that in the analysis of a comparison (*any* comparison), the conformance probability p_c be calculated using Eqs. (9) and (8) (when is reasonable to assume underlying normality). A minimum threshold p_{cL} should be established for p_c and the acceptance criterion should be $p_c \geq p_{cL}$. If the criterion is not met, *i.e.* if $p_c < p_{cL}$, U_x, the uncertainty of the CMC claim, should be adjusted in such a way that p_c reaches the threshold.

There remains to decide a suitable threshold. The simple-acceptance decision rule, described in Section 5.5, sets $p_{cL} = 50\,\%$. Also, the most commonly adopted corrective practice, to enlarge U_x so that $U_x = \Delta x$, follows essentially this decision rule.

Setting $p_{cL} = 50\,\%$ would represent a considerable progress with respect to the current situation, in which a result can be accepted with a conformance probability as low as 20.4 %.

However, we argue that the threshold $p_{cL} = 50\,\%$ is still too low or, to say better, that the consumer's risk is too high. One should be aware that $p_c = 50\,\%$ means that there is a 50 % chance of being incorrect, whatever the decision, acceptance or rejection. As a consequence, there is a 50 % risk that measurement results produced by the accepted laboratory are not traceable[5] to the corresponding units, an uncomfortable situation in our opinion, for which corrective measures should be considered. A possible action would be to raise the value of p_{cL}, thus adopting a guarded acceptance.

Table 1
Worst-case conformance probabilities ($|E_n| \approx 1$); u: arbitrary unit.

$u(x_{ref})/u_x$	$\rho = 0$		$\rho = u(x_{ref})/u_x$	
	Δx_{max}/u	p_c/%	Δx_{max}/u	p_c/%
≈0	$\approx U_x$	≈50	$\approx U_x$	≈50
1/3	$U_x + 0.16U(x_{ref})$	37	$U_x - 0.17U(x_{ref})$	63
0.8	$U_x + 0.35U(x_{ref})$	24	$U_x - 0.5U(x_{ref})$	84
≈1	$\approx\sqrt{2}U_x$	≈20	≈0	≈100

6. Example

We calculated the conformance probabilities p_c in a real case, the Key Comparison APMP.L-K4 [51], which relates to diameter measurement conducted in 2008–2010 with $n = 14$ participating laboratories. The same comparison is analysed in [9]. In this comparison, the KCRV was estimated as the weighted mean of the participants' results, after removing the data from laboratories 2, 7 and 8. The KCRV was $x_{ref} = 0.459\,\mu m$ with $u(x_{ref}) = 0.027\,\mu m$. The small uncertainty of the KCRV compared to those of the participants is typical of a well-behaving KC.

Table 2 shows the reported deviations x_i from the nominal value of 11.95 mm of the internal diameter of a ring and the associated standard uncertainties $u(x_i)$, together with the deviations from the KCRV, Δx_i, and the associated uncertainties $u(\Delta x_i)$. The last two columns give the normalised errors and the conformance probabilities calculated as explained in Section 5.5. As expected, the conformance probabilities confirm generally the current acceptance criterion based on the normalised error, yet some surprises arise when E_n is close to 1. For example, laboratory 6 has $E_n = 0.9$, yet, $p_c = 99.8\,\%$, a very high value. So, a claim (anyway accepted), yielding a comparatively poor performance if evaluated using the normalised error, does indeed have a very low consumer's risk. On the contrary, laboratory 12, which is also the one with the lowest uncertainty, has $E_n = 1.1$, so that the result would not be accepted. However, $p_c = 68\,\%$, a value higher than the current threshold. We think that rejecting a result with $p_c = 68\,\%$, and accepting results with $p_c < 50\,\%$, as currently happens, is unfair. The result should be accepted, and it would be accepted should the method we propose be adopted.

This simple example highlights the perverse consequences of the uncontrolled guard bands introduced by the current validation criterion.

In conclusion, the decision based on p_c is more reliable than that based on E_n, because it is more closely linked to the effective capabilities of the laboratory, as demonstrated when E_n is close to 1.

[5] We are aware that traceability, as currently defined (see [12], definition 2.41), is a yes/no property. We adopt here a broader view of the term. For a motivation of this choice, see [50].

Table 2
Data, normalised errors and conformance probabilities for the Key Comparison APMP.L-K4.

Lab	x_i/μm	$u(x_i)$/μm	Δx_i/μm	$U(\Delta x_i)$/μm	E_n	p_c/%
1	0.43	0.133	−0.029	0.260	−0.1	100
2	0.16	0.0875	−0.299	0.183	−1.7	0
3	0.50	0.30	0.041	0.598	0.1	100
4	0.43	0.087	−0.029	0.165	−0.2	100
5	0.45	0.066	−0.009	0.120	−0.1	100
6	0.00	0.27	−0.459	0.537	−0.9	100
7	−0.30	0.22	−0.759	0.433	−1.7	0
8	−0.99	0.144	−1.449	0.293	−5.0	0
9	0.23	0.28	−0.229	0.557	−0.4	100
10	0.27	0.075	−0.189	0.140	−1.2	7
11	0.35	0.177	−0.109	0.350	−0.3	100
12	0.54	0.047	0.081	0.077	1.1	68
13	0.53	0.064	0.071	0.116	0.6	98
14	0.24	0.58	−0.219	1.159	−0.2	100

7. CMCs in mass metrology

The landmark revision of the SI that came into force on 19 May 2020 implies that in principle every laboratory capable of realising the kilogram in terms of the Planck constant can disseminate the unit on its own. In practice, some experimental results suggested to postpone this situation until a better agreement among the various realisations is achieved, in order to preserve the very good worldwide harmonisation of mass metrology [52]. The Consultative committee for mass and related quantities, CCM, established a strategy for the dissemination of the unit [53], largely based on a previous paper [54]. The key points relevant here are:

- The unit of mass acquires an uncertainty component due to the realisation of the unit, its value at the moment of writing [55] being $u_{CV} = 20\,\mu g$ (where CV stands for the Consensus Value from which the traceability chain starts);
- Accordingly, all the 1 kg mass standards inherit this additional component of uncertainty, as well as all multiples and sub-multiples of the kilogram, suitably scaled.

As the typical uncertainties in the comparison of 1 kg standards are much smaller than this additional component, the future mass comparisons will be characterised by a very high correlation among the results of all the participants.

As an example we consider a fictitious Key Comparison for a 1 kg mass standard with $n = 6$ participating laboratories. Table 3 (columns 2 and 3) and Fig. 1 show the reported deviations x_i from the nominal value of 1 kg and the associated expanded uncertainties. With these data, and a common covariance $u_{i,j} = 400\,\mu g^2$ $(i, j = 1, \ldots, 6,\ i \neq j)$, we obtain from Eq. (3) $\chi^2_{obs} = 22.2$, far exceeding the limit discussed in Section 3, in this case $q_{0.95} = 11.1$.

We thus derive the KCRV x_{ref} as the WM of the largest consistent subset [56], in this case intuitively the set of five results after removal of result 6, with $\chi^2_{obs} = 9.48 \approx q_{0.95}$, obtaining $x_{ref} = -0.12\,\mu g$ with $u(x_{ref}) = 21.42\,\mu g$, $U(x_{ref}) = 42.84\,\mu g$, the green horizontal lines in Fig. 1 representing the corresponding coverage interval. We then calculate $\Delta x_i = x_i - x_{ref}$ and $U(\Delta x_i)$ from $u^2(\Delta x_i) = u^2(x_i) + u^2(x_{ref}) - 2u(x_i, x_{ref})$ (columns 3 and 4 in Table 3). The covariance term for $i = 1, \ldots, 5$ is, as it is well-known, $u(x_i, x_{ref}) = u^2(x_{ref})$, so that $u^2(\Delta x_i) = u^2(x_i) - u^2(x_{ref})$. As to laboratory 6, a superficial reasoning (which we made in a first instance) would conclude that $u(x_6, x_{ref}) = 0$, on the argument that x_{ref} and x_6, the latter being excluded from the calculation of the former, should be uncorrelated. It can be demonstrated that (in agreement with common sense) x_6 and x_{ref} are indeed correlated, their covariance being $u(x_6, x_{ref}) = u^2_{CV} = 400\,\mu g^2$.

Columns 5 and 6 give normalised error and conformance probability, respectively. Surprisingly, according to the former, laboratory 1 turns out not to be consistent. This result is in striking contrast with

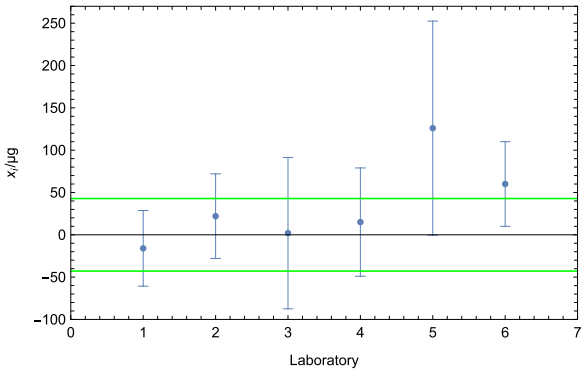

Fig. 1. A fictitious mass comparison. Error bars indicate expanded uncertainties and green lines a coverage interval $[-U(x_{ref}), U(x_{ref})]$ for $x_{ref} = -0.1$.

Table 3
Example of mass comparison.

Lab	x/μg	U_x/μg	Δx/μg	$U(\Delta x)$/μg	E_n	p_c/%
1	−16.0	44.7	−15.9	12.8	−1.24	91
2	22.0	50.0	22.1	25.8	0.86	90
3	2.0	89.4	2.1	78.5	0.03	100
4	15.0	64.0	15.1	47.6	0.32	99
5	126.0	126.5	126.1	119.0	1.06	50
6	60.0	50.0	60.1	33.7	1.78	32

a simple visual inspection of Fig. 1, and is a perverse consequence of the large covariance. Note that assuming $u(x_6, x_{ref}) = 0$ would lead to $U(\Delta x_6) = 66\,\mu g$ and $E_n = 0.91$, so that the claim of laboratory 6 would be accepted. The considerations above represent a word of warning as concerns the correct calculation of E_n in future mass comparisons. On the contrary, the conformance probability, calculated according to Eq. (8), correctly (and easily) captures the real situation.

8. Conclusions

We think that the score $|E_n| \leq 1$ is too weak in any case in a consistency test, for the reasons discussed in Section 4. Furthermore, it is unacceptably weak in the evaluation of the performance of a candidate laboratory in a KC in view of the possible acceptance of a CMC.

In addition, behind that seemingly unique decision rule, different rules – guarded acceptance, simple acceptance and guarded rejection – come into play depending on the specific comparison.

Unpleasant consequences of the facts expounded above are:

- CMC claims are accepted with a conformance probability that can be well below $p_c = 50\,\%$ (see Section 5.5), and
- claims having a far higher conformance probability can be rejected (see Section 6).

The former fact implies that it is more likely to be wrong than right in accepting the claim.

The latter fact is unfair and simply unacceptable.

We argue that the key parameter in the decision whether to accept or not a CMC claim should be the conformance probability, rather than the normalised error. After a check of the consistency of the data set, the acceptance or rejection (and possibly adaptation) of an individual CMC should be decided using conformance probability, which therefore should be calculated using Eq. (8) with the appropriate PDF $g_{\Delta X}(\xi)$ (possibly a normal in most cases).

In this new paradigm, a suitable lower threshold for conformance probability should be agreed at the international level, similarly to the upper threshold so far agreed for the normalised error. Equivalence

should be granted for those laboratories whose claims have a conformance probability greater than the threshold. In the negative, the uncertainty claim should be increased so that the threshold is reached. As concerns the value of the threshold, it is related to the decision rule. We favour a guarded acceptance over simple acceptance, for the reasons discussed in Section 5.6.

Ultimately, the conformance probability, in the case of a CMC claim, is strictly related to the risk of losing traceability for measurement results based on calibrations and measurements made by the laboratory under scrutiny. By adopting the paradigm we proposed here, the risk can be at least known and, ideally, controlled.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are indebted to Francesca Pennecchi (INRiM), Martin Milton (BIPM) and Maurice Cox (NPL) for their critical reading of the manuscript and for the suggestions that led to a significant improvement of the presentation.

Appendix. Normalised error and χ^2

We calculate here the χ^2_{obs} statistic of Eq. (4) for a data set with two data.

In this case, introducing the residuals $r_i = x_i - \hat{\mu}$,

$$\chi^2_{\text{obs}} = \begin{bmatrix} r_1 & r_2 \end{bmatrix} \begin{bmatrix} u_1^2 & u_{1,2} \\ u_{1,2} & u_2^2 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \frac{r_1^2 u_2^2 + r_2^2 u_1^2 - 2 r_1 r_2 u_{1,2}}{u_1^2 u_2^2 - u_{1,2}^2}. \tag{10}$$

Using the weighted mean of the two data

$$\hat{\mu} = \frac{\left(u_2^2 - u_{1,2}\right) x_1 + \left(u_1^2 - u_{1,2}\right) x_2}{u_1^2 + u_2^2 - 2u_{1,2}} \tag{11}$$

as location estimator, the residuals r_i take the form

$$r_1 = \frac{\left(u_i^2 - u_{1,2}\right)\left(x_1 - x_2\right)}{u_1^2 + u_2^2 - 2u_{1,2}} \tag{12}$$

and

$$r_2 = \frac{\left(u_2^2 - u_{1,2}\right)\left(x_2 - x_1\right)}{u_1^2 + u_2^2 - 2u_{1,2}} \tag{13}$$

and, by substituting Eqs. (12) and (13) into Eq. (10)

$$\chi^2_{\text{obs}} = \frac{\left(x_1 - x_2\right)^2}{u_1^2 + u_2^2 - 2u_{1,2}} = 4E_{\text{n}}^2. \tag{14}$$

The χ^2_{obs} is, not too surprisingly (and up to a factor) the square of the normalised error E_{n}, i.e., the quotient of the difference between two estimates and the uncertainty of the difference.

It is worth noting that, *ceteris paribus*, χ^2_{obs} increases as correlation between estimates increases. For $u_{1,2} = u_1^2$, χ^2_{obs} is

$$\chi^2_{\text{obs}} = \frac{\left(x_1 - x_2\right)^2}{u_2^2 - u_1^2}, \tag{15}$$

which diverges fast for $u_2 \to u_1$, unless $x_2 \to x_1$.

References

[1] BIPM, Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, Tech. rep., Bureau International des Poids et Mesures, Sèvres, France, 2003, URL https://www.bipm.org/documents/20126/43742162/CIPM-MRA-2003.pdf.

[2] BIPM, BIPM key comparison database, 2021, URL https://www.bipm.org/kcdb/.

[3] ILAC, The ILAC Mutual Recognition Arrangement, ILAC B7:10/2015, 2015, URL https://ilac.org/?ddownload=891.

[4] BIPM, Measurement comparisons in the CIPM MRA – Guidelines for organizing, participating and reporting, CIPM MRA-G-11, 2021, URL https://www.bipm.org/documents/20126/43742162/CIPM-MRA-G-11.pdf/9fe6fb9a-500c-9995-2911-342f8126226c?version=1.8&download=true.

[5] ISO, Conformity Assessment - General Requirements for Proficiency Testing, International Organization for Standardization (ISO), Geneva, Switzerland, 2010, ISO/IEC 17043:2010.

[6] ISO, Statistical Methods for use in Proficiency Testing by Interlaboratory Comparison, International Organization for Standardization (ISO), Geneva, Switzerland, 2015, ISO 13528:2015.

[7] BIPM, Calibration and measurement capabilities in the context of the CIPM MRA – Guidelines for their review, acceptance and maintenance, CIPM MRA-G-13, 2021, URL https://www.bipm.org/documents/20126/43742162/CIPM-MRA-G-13.pdf/f8b8c429-42e0-4cf1-dc6c-bc60ab7f371a?version=1.5&download=true.

[8] M. Cox, P. Harris, M. Milton, Method for determining acceptable CMCs to ensure consistency with KC results, CCQM Report 09-15, 2009, URL https://www.bipm.org/cc/CCQM/Restricted/15/CCQM-09-15.pdf.

[9] K. Shirono, M. Cox, Statistical reassessment of calibration and measurement capabilities based on key comparison results, Metrologia 56 (4) (2019) 045001, http://dx.doi.org/10.1088/1681-7575/ab219e.

[10] BIPM/ILAC working group, Calibration and measurement capabilities. A paper by the joint BIPM/ILAC working group, Working document CIPM 2007-11, 2007, URL https://www.bipm.org/utils/common/documents/jcrb/CIPM_2007_11_CMC_BMC_accepted.pdf.

[11] ILAC, ILAC Policy for Measurement Uncertainty in Calibration, ILACP14:09/2020, 2020, URL https://ilac.org/?ddownload=123348.

[12] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, International vocabulary of metrology — Basic and general concepts and associated terms, Joint Committee for Guides in Metrology, JCGM 200:2012, 2012, URL http://www.bipm.org/en/publications/guides/vim.html.

[13] P.W. Bridgman, The Logic of Modern Physics, McMillan, 1927, URL https://archive.org/details/logicofmodernphy00brid.

[14] P. Bickel, K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, in: Holden-Day Series in Probability and Statistics, Prentice Hall, 1977, URL https://books.google.com.vc/books?id=I_AuswEACAAJ.

[15] C.R. Rao, Linear Statistical Inference and its Applications, 2nd Ed., John Wiley & Sons, 1973.

[16] ISO, Statistics — Vocabulary and Symbols — Part 1: Probability and General Statistical Terms, International Organization for Standardization (ISO), Geneva, Switzerland, 2006, ISO 3534-1:2006.

[17] B. van der Waerden, V. Thompson, E. Sherman, Mathematical Statistics, in: Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2013, URL https://books.google.it/books?id=zLztCAAAQBAJ.

[18] W. Rand, Introduction to Robust Estimation and Hypothesis Testing, Academic Press, San Diego, 1997, http://dx.doi.org/10.2307/2669876.

[19] M.G. Cox, The evaluation of key comparison data: An introduction, Metrologia 39 (6) (2002) 587–588, http://dx.doi.org/10.1088/0026-1394/39/6/9.

[20] M.G. Cox, The evaluation of key comparison data, Metrologia 39 (6) (2002) 589–595, http://dx.doi.org/10.1088/0026-1394/39/6/10.

[21] C. Woolston, Psychology journal bans *P* values, Nature 519 (7541) (2015) 9, http://dx.doi.org/10.1038/519009f.

[22] M. Baker, Statisticians issue warning over misuse of *P* values, Nature 531 (7593) (2016) 151, http://dx.doi.org/10.1038/nature.2016.19503.

[23] D. Singh Chawla, Big names in statistics want to shake up much-maligned *P* value, Nature 548 (7665) (2017) 16–17, http://dx.doi.org/10.1038/nature.2017.22375.

[24] K. Kafadar, EDITORIAL: Statistical significance, *P*-values, and replicability, Annals Appl. Stat. 15 (3) (2021) 1081–1083, http://dx.doi.org/10.1214/21-AOAS1500.

[25] Y. Benjamini, R.D.D. Veaux, B. Efron, S. Evans, M. Glickman, B.I. Graubard, X. He, X.-L. Meng, N. Reid, S.M. Stigler, S.B. Vardeman, C.K. Wikle, T. Wright, L.J. Young, K. Kafadar, The ASA president's task force statement on statistical significance and replicability, Annals Appl. Stat. 15 (3) (2021) 1084–1085, http://dx.doi.org/10.1214/21-AOAS1501.

[26] P.J. Mohr, B.N. Taylor, CODATA recommended values of the fundamental physical constants: 1998, J. Phys. Chem. Refer. Data 28 (6) (1999) 1713–1852, http://dx.doi.org/10.1063/1.556049, arXiv:https://doi.org/10.1063/1.556049.

[27] R.T. Birge, Probable values of the general physical constants, Rev. Modern Phys. 1 (1929) 1–73, http://dx.doi.org/10.1103/RevModPhys.1.1, URL https://link.aps.org/doi/10.1103/RevModPhys.1.1.

[28] O. Bodnar, C. Elster, On the adjustment of inconsistent data using the Birge ratio, Metrologia 51 (5) (2014) 516–521, http://dx.doi.org/10.1088/0026-1394/51/5/516.

[29] O. Bodnar, C. Elster, J. Fischer, A. Possolo, B. Toman, Evaluation of uncertainty in the adjustment of fundamental constants, Metrologia 53 (1) (2016) S46–S54, http://dx.doi.org/10.1088/0026-1394/53/1/s46.

[30] C. Merkatas, B. Toman, A. Possolo, S. Schlamminger, Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation, Metrologia 56 (5) (2019) 054001, http://dx.doi.org/10.1088/1681-7575/ab3365.

[31] K. Weise, W. Wöger, Removing model and data non-conformity in measurement evaluation, Meas. Sci. Technol. 11 (12) (2000) 1649–1658, http://dx.doi.org/10.1088/0957-0233/11/12/301.

[32] C. Elster, B. Toman, Analysis of key comparison data: critical assessment of elements of current practice with suggested improvements, Metrologia 50 (5) (2013) 549–555, http://dx.doi.org/10.1088/0026-1394/50/5/549.

[33] O. Bodnar, C. Elster, Analysis of key comparisons with two reference standards: Extended random effects meta-analysis, in: A.B. Forbes, N.-F. Zhang, A. Chunovkina, S. Eichstädt, F. Pavese (Eds.), Advanced Mathematical and Computational Tools in Metrology and Testing XI, WORLD SCIENTIFIC, 2018, pp. 1–8, http://dx.doi.org/10.1142/9789813274303_0001, arXiv:https://www.worldscientific.com/doi/pdf/10.1142/9789813274303_0001, URL https://www.worldscientific.com/doi/abs/10.1142/9789813274303_0001.

[34] W.E. Strawderman, A.L. Rukhin, Simultaneous estimation and reduction of nonconformity in interlaboratory studies, J. R. Stat. Soc. Ser. B 72 (2) (2010) 219–234, http://dx.doi.org/10.1111/j.1467-9868.2009.00733.x, arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2009.00733.x, URL https://www.worldscientific.com/doi/abs/10.1142/9789813274303_0001.

[35] R. DerSimonian, N. Laird, Meta-analysis in clinical trials, Control. Clin. Trials 7 (3) (1986) 177–188, http://dx.doi.org/10.1016/0197-2456(86)90046-2, URL https://www.sciencedirect.com/science/article/pii/0197245686900462.

[36] R. DerSimonian, N. Laird, Meta-analysis in clinical trials revisited, Contemp. Clin. Trials 45 (Pt A) (2015) 139–145.

[37] R.C. Paule, J. Mandel, Consensus values, regressions, and weighting factors, J. Res. Nat. Inst. Stand. Technol. 94 (3) (1989) 197–203, http://dx.doi.org/10.6028/jres.094.020, URL https://pubmed.ncbi.nlm.nih.gov/28053410.

[38] R.C. Paule, J. Mandel, Consensus values and weighting factors, J. Res. Nat. Bureau Stand. 87 (5) (1982).

[39] A. Koepke, T. Lafarge, B. Toman, A. Possolo, NIST Consensus Builder — User's Manual, National Institute of Standards and Technology, Gaithersburg, MD, 2017, URL https://consensus.nist.gov.

[40] A. Koepke, T. Lafarge, A. Possolo, B. Toman, Consensus building for interlaboratory studies, key comparisons, and meta-analysis, Metrologia 54 (3) (2017) S34–S62, http://dx.doi.org/10.1088/1681-7575/aa6c0e.

[41] M. Thompson, S. Ellison, R. Wood, The international harmonized protocol for the proficiency testing of analytical chemistry laboratories: (IUPAC technical report), Pure Appl. Chem. Pure Appl. Chem. 78 (2006) 145–196, http://dx.doi.org/10.1351/pac200678010145.

[42] W. Wöger, Remarks on the E_n-criterion used in measurement comparisons, PTB-Mitteilungen 109 (1) (1999) 24–27, URL https://serials.unibo.it/cgi-ser/start/en/spogli/df-s.tcl?prog_art=5933020&language=ENGLISH&view=articoli.

[43] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, Evaluation of measurement data — The role of measurement uncertainty in conformity assessment, Joint Committee for Guides in Metrology, JCGM 106:2012, 2012.

[44] D. Sivia, Data Analysis: A Bayesian Tutorial, in: Data Analysis: A Bayesian Tutorial, Clarendon Press, 1996, URL https://books.google.it/books?id=wR5yljKasLsC.

[45] P. Lee, Bayesian Statistics: An Introduction, Edward Arnold, 1989.

[46] G. Wübbeler, O. Bodnar, B. Mickan, C. Elster, Explanatory power of degrees of equivalence in the presence of a random instability of the common measurand, Metrologia 52 (2) (2015) 400–405, http://dx.doi.org/10.1088/0026-1394/52/2/400.

[47] J. Wright, B. Toman, B. Mickan, G. Wübbeler, O. Bodnar, C. Elster, Transfer standard uncertainty can cause inconclusive inter-laboratory comparisons, Metrologia 53 (6) (2016) 1243–1258, http://dx.doi.org/10.1088/0026-1394/53/6/1243.

[48] Laboratoire national de métrologie et d'essais, CaSoft: Software for conformity assessment taking into account measurement uncertainty - version 2, 2019, URL https://www.lne.fr/en/software/CASoft.

[49] A. Allard, N. Fischer, I. Smith, P. Harris, L. Pendrill, Risk calculations for conformity assessment in practice, in: Array (Ed.), International Congress of Metrology, 2019, p. 16001, http://dx.doi.org/10.1051/metrology/201916001.

[50] W. Bich, Interdependence between measurement uncertainty and metrological traceability, Accreditation Qual. Assurance: J. Qual. Compar. Reliab. Chem. Measur. 14 (2009) 581–586, http://dx.doi.org/10.1007/s00769-009-0500-4.

[51] J.-H. Chin, T. Takatsuji, M. Horita, T. Hamakawa, K.P. Chaudhary, A. Tonmueanwai, N. Alfiyati, O. Kruger, E. Howick, P. Cox, M. bin Sawi, B.Q. Thu, J.-A. Kim, W.S. Yin, T.S. Leng, S.A. Zaher, Final report on key comparison APMP.L-K4: Calibration of diameter standards, Metrologia 51 (1A) (2014) 04004, http://dx.doi.org/10.1088/0026-1394/51/1a/04004.

[52] M. Gläser, M. Borys, D. Ratschko, R. Schwartz, Redefinition of the kilogram and the impact on its future dissemination, Metrologia 47 (4) (2010) 419, URL http://stacks.iop.org/0026-1394/47/i=4/a=007.

[53] CCM Consultative committee for mass and related quantities, CCM Detailed Note on the Dissemination Process after the Redefinition of the kilogram, CCM/2019-06B, 2019, URL https://www.bipm.org/documents/20126/28432674/working-document-ID-11291/cf8f685d-fc3d-1883-9a42-0678a2c34453.

[54] M. Stock, S. Davidson, H. Fang, M. Milton, E. de Mirandés, P. Richard, C. Sutton, Maintaining and disseminating the kilogram following its redefinition, Metrologia 54 (6) (2017) S99, URL http://stacks.iop.org/0026-1394/54/i=6/a=S99.

[55] S. Davidson, M. Stock, Beginning of a new phase of the dissemination of the kilogram, Metrologia 58 (3) (2021) 033002, http://dx.doi.org/10.1088/1681-7575/abef9f.

[56] M.G. Cox, The evaluation of key comparison data: determining the largest consistent subset, Metrologia 44 (3) (2007) 187–200, http://dx.doi.org/10.1088/0026-1394/44/3/005.