Metrologia **41** (2004) 340–352 PII: S0026-1394(04)81455-5

# Statistical analysis of CIPM key comparisons based on the ISO *Guide*

**R N Kacker, R U Datla and A C Parr**

National Institute of Standards and Technology, Gaithersburg, MD 20899-8910, USA

E-mail: raghu.kacker@nist.gov

Received 13 April 2004
Published 30 June 2004
Online at stacks.iop.org/Met/41/340
doi:10.1088/0026-1394/41/4/017

**Abstract**

An international Advisory Group on Uncertainties has published guidelines for the statistical analysis of a simple key comparison carried out by the Consultative Committees of the International Committee of Weights and Measures (CIPM) where a travelling standard of a stable value is circulated among the participants. We discuss several concerns regarding these guidelines. Then, we describe a statistical model based on the *Guide to the Expression of Uncertainty in Measurement* to establish the key comparison reference value, the degrees of equivalence, and their associated standard uncertainties on the basis of the data submitted by the participants. The proposed statistical model applies to all those CIPM key comparisons where the submitted results are mutually comparable and appropriate for determining the key comparison reference value and the submitted uncertainties are sufficiently reliable.

## 1. Introduction

Key comparisons are interlaboratory comparisons between national metrology institutes (NMIs) [1]. Key comparisons carried out by the Consultative Committees (CCs) of the Comité International des Poids et Mesures (International Committee of Weights and Measures, CIPM) or the Bureau International des Poids et Mesures (BIPM)[1] are referred to as CIPM key comparisons. The outputs of the statistical analysis of a CIPM key comparison are the key comparison reference value, the degrees of equivalence, and their associated uncertainties [1]. An international Advisory Group on Uncertainties[2] commissioned by the director of the BIPM has published guidelines for the statistical analysis of a CIPM key comparison [2]. We refer to these guidelines as the Statistical Guidelines paper to distinguish it from the previously published Guidelines for CIPM key comparisons [3]. The Statistical Guidelines paper applies to a simple CIPM key comparison where a travelling standard having good short-term stability and stability during transport is circulated among the participating NMI laboratories and each NMI laboratory realizes its measurement independent of the others. The Statistical Guidelines paper notes that complications such as the following may occur: some or all of the measurements are mutually dependent, the travelling standard is not stable, the pattern of comparison is complicated, the reference value is provided in advance by some means, several travelling standards are circulated, or the measurements are made at various settings of a parameter such as wavelength or frequency. The Advisory Group on Uncertainties intends to develop further guidelines to cover these and other complications. Many CIPM key comparisons are not simple. So the Statistical Guidelines paper does not apply to many CIPM key comparisons.

The Statistical Guidelines paper consists of two statistical procedures, A and B. Procedure A applies to consistent laboratory results; it is based on frequentist statistics. Procedure B applies when the laboratory results are inconsistent; it is based on the concept of a *measurement equation* introduced by the International Organization for Standardization (ISO) *Guide to the Expression of Uncertainty in Measurement* [4]. Procedure A begins with a chi-squared test to assess the consistency of the results. When the results are judged to be consistent, the Statistical Guidelines paper recommends continuing with procedure A. When the results are judged to be inconsistent, the NMI laboratories

[1] The BIPM operates under the exclusive supervision of the CIPM. ‘The CIPM is the world’s highest authority in the field of measurement science (i.e. metrology),’ (http://physics.nist.gov/cuu/Uncertainty/international1.html).

[2] The Advisory Group consisted of five experts from the NMIs of Denmark, Germany, Italy, UK, and USA.

0026-1394/04/040340+13$30.00

with inconsistent results are given a chance to self-investigate the inconsistent results and given the option of withdrawing from the comparison, subject to the protocol and limitations of the available time and resources. When the inconsistent results are not eliminated, the Statistical Guidelines paper recommends procedure B. Procedure A recommends as the key comparison reference value the weighted mean of laboratory results where the weights are inversely proportional to the squares of submitted standard uncertainties (variances). Procedure B recommends as the key comparison reference value an expected median determined through numerical simulation.

The Statistical Guidelines paper does not state the statistical models that underlie its procedures A and B and it does not discuss statistical interpretation of the outputs of these procedures. We note the following concerns regarding the Statistical Guidelines paper. The expected values of the sampling distributions of all degrees of equivalence, whether small or large, determined from procedure A are zero; therefore, they do not quantitate the agreements and disagreements between the results, defeating their purpose. The chi-squared test does not justify the frequentist statistics model that underlies the use of the weighted mean as the key comparison reference value. The uncertainties associated with the key comparison reference values, determined using procedures A and B, are underestimates because the corresponding measurement equations are incomplete.

This paper describes a systematic laboratory effects model for the statistical analysis of CIPM key comparisons based on the ISO *Guide*. Sections 2 and 3 are, respectively, a review and a criticism of the Statistical Guidelines paper. In section 4, we present a brief description of the systematic laboratory effects model proposed in [5]. Then in section 5, we outline a statistical analysis to determine the key comparison reference value and its associated uncertainty based on the systematic laboratory effects model. In section 6, we discuss a useful statistical method for identifying discrepant results and uncertainties. In section 7, we illustrate the proposed statistical analysis using the data from a supplementary comparison of cryogenic radiometers. A summary appears in section 8.

## 2. Review of the Statistical Guidelines paper

The Statistical Guidelines paper applies to a simple CIPM key comparison where a travelling standard serves as the common measurand for all participating NMI laboratories and the results are statistically independent. The value of the travelling standard is believed to be constant during the comparison. We use the symbol $Y$ for the value of the common measurand. The data from a CIPM key comparison consist of the paired results and standard uncertainties $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ submitted by the participating NMI laboratories[3]. The results $x_1, \ldots, x_n$ relate to the national measurement standards maintained by the NMI laboratories. The outputs of a statistical analysis of the data from a CIPM key comparison are the key comparison reference value $x_R$, the degree of equivalence $d_i = x_i - x_R$ of the result $x_i$, the degree of equivalence $d_{i,j} = x_i - x_j$ of $x_i$ and $x_j$, and their associated standard uncertainties $u(x_R)$, $u(d_i)$, and $u(d_{i,j})$, respectively, for $i, j = 1, 2, \ldots, n$ and $i \neq j$ [1]. We refer to the results $x_1, \ldots, x_n$ as laboratory results. We use the symbols $X_1, \ldots, X_n$ for the expected values, $E(x_1), \ldots, E(x_n)$, of the sampling distributions of $x_1, \ldots, x_n$, respectively[4]. We refer to the expected values $X_1, \ldots, X_n$ as the laboratory expected values. We use the symbols $\sigma_1, \ldots, \sigma_n$ for the true standard deviations, $S(x_1), \ldots, S(x_n)$, of the sampling distributions of $x_1, \ldots, x_n$, respectively. The uncertainties $u(x_1), \ldots, u(x_n)$ are estimates of $\sigma_1, \ldots, \sigma_n$, respectively.

Procedure A recommends the weighted mean $x_W = \sum_i w_i x_i / \sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, as the key comparison reference value, $x_R$, and the expression $u(x_W) = 1/\sqrt{[\sum_i w_i]}$ as $u(x_R)$. This recommendation is based on the assumption that the submitted variances, $u^2(x_1), \ldots, u^2(x_n)$, are equal to the true variances, $\sigma_1^2, \ldots, \sigma_n^2$, of the sampling distributions of $x_1, \ldots, x_n$, respectively. Procedure A recommends the following expressions for the degrees of equivalence and their associated uncertainties: $d_i = x_i - x_W$ and $u(d_i) = \sqrt{[u^2(x_i) - u^2(x_W)]}$, $d_{i,j} = x_i - x_j$ and $u(d_{i,j}) = \sqrt{[u^2(x_i) + u^2(x_j)]}$, for $i$, $j = 1, 2, \ldots, n$ and $i \neq j$.

Procedure B recommends a simulated expected median, $x_M$, and an uncertainty, $u(x_M)$, determined through numerical simulation as the key comparison reference value, $x_R$, and uncertainty, $u(x_R)$, respectively. Procedure B is as follows. The laboratory expected values, $X_1, \ldots, X_n$, are regarded as random variables having state-of-knowledge probability distributions with expected values $x_1, \ldots, x_n$ and standard deviations $u(x_1), \ldots, u(x_n)$, respectively. Procedure B assumes that the joint probability distribution of the vector $(X_1, \ldots, X_n)$ is multivariate normal (Gaussian) or some other fully specified distribution. Generate one million ($10^6$) simulated random sample vectors $(x_1^{(r)}, \ldots, x_n^{(r)})$ from the assumed joint probability distribution for $(X_1, \ldots, X_n)$, where $r$ is an index for the sample number. Calculate the median $m^{(r)} = \text{median}(x_1^{(r)}, \ldots, x_n^{(r)})$ for each random sample vector. Then $x_M$ is the arithmetic mean and $u(x_M)$ is the standard deviation of the one million simulated medians, $m^{(r)}$. The degrees of equivalence $d_i$ and $d_{i,j}$ are $d_i = x_i - x_M$ and $d_{i,j} = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The corresponding uncertainties, $u(d_i)$ and $u(d_{i,j})$, are calculated from the one million simulated random sample vectors $(x_1^{(r)}, \ldots, x_n^{(r)})$.

## 3. Criticism of the Statistical Guidelines paper

Here are our concerns regarding the Statistical Guidelines paper.

### *3.1. Interpretation of the degrees of equivalence*

Procedure A is based on the following model from frequentist statistics. The results are regarded as realizations of the random variables $x_1, \ldots, x_n$, where

$$x_i = Y + e_i \qquad (1)$$

[3] The result $x_i$ includes corrections for recognized systematic effects applied in the laboratory labelled $i$ and the uncertainty $u(x_i)$ includes the uncertainties associated with the corrections for $i = 1, 2, \ldots, n$.

[4] We use the same symbols, $x_1, \ldots, x_n$, for both the random variables having sampling distributions and the results that are regarded as realizations of the random variables.

and the sampling distributions of $e_1, \ldots, e_n$ are assumed to be mutually independent and normal with expected values zero and variances $u^2(x_1), \ldots, u^2(x_n)$, respectively. Thus the expected value, $E(x_i)$, is equal to $Y$ and the variance, $V(x_i)$, is equal to $u^2(x_i)$, for $i = 1, 2, \ldots, n$. Reference [5] refers to model (1) as a non-existent laboratory effects model to distinguish it from those statistical models that allow for the possibility of laboratory effects (biases) in the results $x_1, \ldots, x_n$. The least-squares estimate of the parameter $Y$ of the non-existent laboratory effects model (1) is the weighted mean $x_W = \sum_i w_i x_i / \sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. According to this model the expected value and the standard deviation[5] of the sampling distribution of $x_W$ are $Y$ and $u(x_W) = 1/\sqrt{[\sum_i w_i]}$, respectively. Thus, the key comparison reference value, $x_R$, based on model (1) is $x_W$ and $u(x_R)$ is $u(x_W)$. The corresponding degrees of equivalence are $d_i = x_i - x_W$ and $d_{i,j} = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. According to model (1), the uncertainty $u(d_i)$ is $\sqrt{[u^2(x_i) - u^2(x_W)]}$ and the uncertainty $u(d_{i,j})$ is $\sqrt{[u^2(x_i) + u^2(x_j)]}$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. In the uncertainty $u(d_i)$, the sign is negative because of the covariance[6] between $x_i$ and $x_W$.

In practice, the standard uncertainties, $u(x_1), \ldots, u(x_n)$, submitted by the participating NMI laboratories are statistical estimates of the true standard deviations, $\sigma_1, \ldots, \sigma_n$, respectively. So the uncertainties $u(x_1), \ldots, u(x_n)$ are uncertain ([4], section E.4). The expression $1/\sqrt{[\sum_i w_i]}$ does not include a component of uncertainty for the uncertainties in the submitted uncertainties, $u(x_1), \ldots, u(x_n)$. Therefore, $u(x_W) = 1/\sqrt{[\sum_i w_i]}$ is an underestimate of the true standard deviation of the weighted mean, $x_W$ [6, 7].

Let us discuss the statistical interpretation of the key comparison reference value determined from the non-existent laboratory effects model (1). According to the non-existent laboratory effects model, the key comparison reference value, $x_R$, is a realization of a random variable[7] with a sampling distribution having the expected value $Y$ and standard deviation $u(x_R) = u(x_W) = 1/\sqrt{[\sum_i w_i]}$. The interval $[x_R \pm 2u(x_R)]$ determined using the non-existent laboratory effects model is a confidence interval for $Y$ computed from the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$. Its coverage property is expressed as a confidence level. The confidence level is not a statement about the computed interval $[x_R \pm 2u(x_R)]$, which either includes or does not include the unknown value, $Y$, of the measurand. It is a statement about the statistical procedure used to compute the interval $[x_R \pm 2u(x_R)]$. Imagine that the CIPM key comparison could be repeated infinitely many times under exactly the same conditions using exactly the same instruments and artefacts. Now imagine that throughout these repetitions exactly the same sampling distributions continued to apply to the random variables $x_1, \ldots, x_n$. The confidence level is the fraction of the infinitely many hypothetical intervals, such as $[x_R \pm 2u(x_R)]$, that would include $Y$ [8]. The ISO *Guide* interprets $x_R$ and $u(x_R)$ as the expected value and standard deviation of a state-of-knowledge distribution for $Y$. The coverage probability of the interval $[x_R \pm 2u(x_R)]$ is the fraction of a state-of-knowledge distribution represented by $x_R$ and $u(x_R)$ that is encompassed by this interval ([4], section 6.2.2). Therefore, the statistical interpretation of $x_R, u(x_R)$, and the interval $[x_R \pm 2u(x_R)]$ based on the non-existent laboratory effects model (1) does not agree with the ISO *Guide*.

Now let us discuss the statistical interpretation of the degrees of equivalence determined from the non-existent laboratory effects model (1). According to the non-existent laboratory effects model, the expected values of the sampling distributions of $x_1, \ldots, x_n$, and $x_R$ are all equal to $Y$. Therefore, the expected values of the sampling distributions of all degrees of equivalence $d_i = x_i - x_R$ and $d_{i,j} = x_i - x_j$ are zero, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. This implies that all computed degrees of equivalence, whether small or large, are statistical estimates of zero. Therefore, the degrees of equivalence determined from the non-existent laboratory effects model (1) do not quantitate the agreements and disagreements between the results, defeating their purpose. In particular, according to this model, all degrees of equivalence published in the key comparison database (KCDB) [9] are estimates of zero.

### 3.2. Limitation of the chi-squared test

The chi-squared test checks whether the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ fit the frequentist statistics model $x_i = \mu + e_i$, for $i = 1, 2, \ldots, n$, where (i) the parameter $\mu$ is any unknown constant, (ii) the errors, $e_1, \ldots, e_n$, are random variables with sampling distributions, (iii) the sampling distributions of $e_1, \ldots, e_n$ are mutually independent and normal (Gaussian) with expected values zero and standard deviations $\sigma_1, \ldots, \sigma_n$, respectively, and (iv) the submitted standard uncertainties, $u(x_1), \ldots, u(x_n)$, are assumed to be the true standard deviations, $\sigma_1, \ldots, \sigma_n$, respectively [10].

The chi-squared test requires that all submitted uncertainties, $u(x_1), \ldots, u(x_n)$, must be sufficiently reliable[8] to be regarded as equal to the true standard deviations, $\sigma_1, \ldots, \sigma_n$, of the sampling distributions of $x_1, \ldots, x_n$. When this requirement is not met, the chi-squared test is not justified.

When the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ reasonably fit the model that underlies the chi-squared test, we say that the results are consistent in view of the submitted uncertainties, $u(x_1), \ldots, u(x_n)$. That is, the dispersion of the results $x_1, \ldots, x_n$ is not more than what can reasonably be attributed to the uncertainties $u(x_1), \ldots, u(x_n)$. When the data do not reasonably fit the model, we say that the results are inconsistent. That is, the dispersion of $x_1, \ldots, x_n$ is more than what can reasonably be attributed to $u(x_1), \ldots, u(x_n)$.

The non-existent laboratory effects model, which underlies the use of the weighted mean as the key comparison reference value, represents the following two-part assumption: the laboratory expected values, $X_1, \ldots, X_n$, are all equal and the common value of $X_1, \ldots, X_n$ is equal to $Y$. When

[5] $E(x_W) = \sum_i w_i E(x_i)/\sum_i w_i = Y(\sum_i w_i/\sum_i w_i) = Y$. If $w_i = 1/u^2(x_i) = 1/V(x_i)$ for $i = 1, 2, \ldots, n$, then the variance of $x_W$ is $V(x_W) = \sum_i w_i^2 V(x_i)/(\sum_i w_i)^2 = \sum_i w_i/(\sum_i w_i)^2 = 1/\sum_i w_i$.

[6] The covariance between $x_i$ and $x_W$ is $C(x_i, x_W) = w_i V(x_i)/\sum_i w_i = 1/\sum_i w_i = V(x_W)$. Therefore, the variance of $d_i$ is $V(x_i - x_W) = V(x_i) + V(x_W) - 2C(x_i, x_W) = V(x_i) + V(x_W) - 2V(x_W) = V(x_i) - V(x_W)$.

[7] The symbol $x_R$ represents both the random variable and its realized value computed from the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$. Likewise, the symbols $d_1, \ldots, d_n$ and $d_{1,2}, \ldots, d_{n-1,n}$ represent both the random variables and their realized values.

[8] In the context of uncertainty in measurement, the adjective 'reliable' refers to the quality of an expression of uncertainty.

the results are judged to be consistent using the chi-squared test, we may say that the data do not refute the assumption that $X_1, \ldots, X_n$ are all equal to some unknown constant $\mu$. The consistency of results does not imply that the unknown common expected value, $\mu$, is equal to the unknown value, $Y$, of the measurand. Therefore, the second part of the assumption is not justified. Thus the chi-squared test does not justify the non-existent laboratory effects model.

The relationship between the result $x_i$ and the unknown value $Y$ of the measurand depends on the relationship between the unknown laboratory expected value, $X_i$, and $Y$, for $i = 1, 2, \ldots, n$. The difference $X_i - Y$ is the unknown (additive) bias[9] in $x_i$ for $i = 1, 2, \ldots, n$. We refer to the bias $X_i - Y$ as the laboratory effect. The chi-squared test does not justify the assumption that the laboratory effects $(X_1 - Y), \ldots,$ $(X_n - Y)$ are all zero.

In the next section, the laboratory effects $(X_1 - Y), \ldots,$ $(X_n - Y)$ are regarded as unknown constants that may be different for different laboratories. The bias in some of the results $x_1, \ldots, x_n$ may be negligible or zero. However, to the extent that the laboratory expected values, $X_1, \ldots, X_n$, and the value $Y$ of the measurand are unknown, it may not be possible to identify such results with certainty.

### 3.3. Underestimation of uncertainty

For the key comparison reference value, $x_R$, and uncertainty, $u(x_R)$, to have a statistical interpretation that agrees with the ISO *Guide*, they must be determined from a measurement equation. All input and output quantities involved in a measurement equation are regarded as variables with state-of-knowledge probability distributions. Following the ISO *Guide*, we use the symbol $Y$ for both the unknown constant value of the measurand and a variable with a state-of-knowledge probability distribution about the value of the measurand. Likewise, we use the symbols $X_1, \ldots, X_n$ for both the unknown laboratory expected values and variables with state-of-knowledge probability distributions about the laboratory expected values. In the ISO *Guide*, the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ are regarded as known constants [4, 8]. The Statistical Guidelines paper assumes that the pairs $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ are the expected values and standard deviations of normal distributions attributed to $Y$ by the participating NMI laboratories. The expected values and/or standard deviations of these distributions are different. So these are different state-of-knowledge distributions for $Y$. Our interpretation is that the result $x_i$ is the expected value, $E(X_i)$, and the standard uncertainty $u(x_i)$ is the standard deviation $S(X_i)$ of a state-of-knowledge distribution for the laboratory expected value $X_i$, for $i = 1, 2, \ldots, n$.

Suppose $x$ is a result of measurement for $Y$ and that its associated standard uncertainty is $u(x)$. Suppose the expected value, $E(x)$, of the sampling distribution of $x$ is $X$. The difference $(X - Y)$ is the bias[10] in $x$. The bias is an unknown constant. Before publication of the ISO *Guide*, there was no generally accepted approach to account for the uncertainty arising from bias. The ISO *Guide* recommends that the result, $x$, should be corrected to counter its possible bias $(X - Y)$ and the uncertainty associated with the correction should be included in the combined standard uncertainty associated with the corrected result. A measurement equation is required to incorporate a correction for possible bias in $x$. The measurement equation that corresponds to the bias $(X - Y)$ is $Y = X + C$, where $C$ is a variable with a probability distribution representing the state-of-knowledge about the expression $(Y - X)$ for the negative bias. In the measurement equation $Y = X + C$, the quantities $X$ and $Y$ are regarded as variables with probability distributions representing states of knowledge about the unknown expected value, $X$, and the unknown value, $Y$, of the measurand. The ISO *Guide* identifies the result, $x$, and uncertainty, $u(x)$, with the expected value, $E(X)$, and standard deviation, $S(X)$, of a state-of-knowledge distribution[11] for $X$, i.e. $E(X) = x$ and $S(X) = u(x)$. A distribution for $C$ is specified independent of the state-of-knowledge distribution for $X$ after the expected value, $x$, and standard deviation, $u(x)$, have been specified. Suppose the expected value, $E(C)$, and standard deviation, $S(C)$, of a state-of-knowledge distribution for $C$ are $c$ and $u(c)$, respectively. Then a corrected combined result (CCR), $y$, for $Y$ is determined by substituting the expected value $x$ for the variable $X$ and the expected value $c$ for the variable $C$ in the measurement equation $Y = X + C$. Thus $y = x + c$. That is, the correction applied to the result $x$ to counter its possible bias is $c$. The combined standard uncertainty, $u(y)$, associated with $y$ is determined by propagating the variances $V(X) = u^2(x)$, $V(C) = u^2(c)$, and the covariance $C(X, C)$. Since the state-of-knowledge distributions for $X$ and $C$ are independent, the covariance $C(X, C)$ is zero; therefore the propagation formula for the measurement equation $Y = X + C$ is $u^2(y) = u^2(x) + u^2(c)$. Thus $u(y) = \surd[u^2(x) + u^2(c)]$. Following the ISO *Guide*, the result $y$ and standard uncertainty $u(y)$ are interpreted as the expected value and standard deviation of a state-of-knowledge distribution for $Y$.

Whenever the bias in a result of measurement with respect to the value of the measurand is unknown, a correction variable should be included in the measurement equation for that measurand. When the bias is believed to be small, the correction variable may have a zero expected value and a small[12] standard deviation. A measurement equation that ignores the correction variable for an unknown bias is incomplete. The uncertainty determined from an incomplete measurement equation is an underestimate.

Let $X_W = \sum_i w_i X_i / \sum_i w_i$ be a linear combination of $X_1, \ldots, X_n$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. When $X_1, \ldots, X_n$ are regarded as variables with state-of-knowledge distributions, $X_W$ is a variable with a state-of-knowledge distribution. The expected value of $X_W$ is $x_W = \sum_i w_i x_i / \sum_i w_i$. When the variables $X_1, \ldots, X_n$ are mutually independent, the standard deviation

[9] We use the word ‘bias’ as a synonym for systematic error. This definition of bias (systematic error) is based on the ISO *Guide* (section B.2.22). In some applications, multiplicative bias, $X_i/Y$, may be more appropriate.

[10] The bias in a result of measurement $x$ is defined with respect to the sampling distribution associated with $x$, whether it is a frequentist or a Bayesian point estimate.

[11] As discussed in [8], this interpretation is justified when $x$ and $u(x)$ are the expected value and standard deviation of a Bayesian posterior distribution for $X$ or are regarded as their approximations.

[12] A zero expected value and a zero standard deviation for the correction variable would imply that the bias is known to be zero. A claim of zero bias may be unjustified. To the extent that the expected value and the value of the measurand are unknown, one cannot be certain that the bias is zero.

of $X_W$ is $u(x_W) = 1/\sqrt{[\sum_i w_i]}$. The measurement equation that corresponds to the use of $x_W$ as $x_R$ and $u(x_W)$ as $u(x_R)$ is $Y = X_W$. This measurement equation is incomplete because it does not include an input variable to represent a correction for possible bias in $x_W$ with respect to $Y$; therefore, $u(x_W) = 1/\sqrt{[\sum_i w_i]}$ is an underestimate of the uncertainty associated with $x_W$. The bias in $x_W$ is the difference $(X_W - Y)$, where $X_W$ is regarded as the expected value of the sampling distribution of $x_W$ and $Y$ is the value of the measurand. To the extent that $X_W$ and $Y$ are unknown, one cannot be certain that the bias in $x_W$ is zero.

The Statistical Guidelines paper does not discuss statistical interpretation of the standard uncertainty, $u(x_R)$, and the uncertainty interval $[x_R \pm 2u(x_R)]$ determined from its procedures A and B. In accordance with the ISO *Guide*, $x_R$ is the expected value and $u(x_R)$ is the standard deviation of a state-of-knowledge distribution for the values that could reasonably be attributed to $Y$. Thus the interval $[x_R \pm 2u(x_R)]$ should include a large fraction, called coverage probability, of the values that could reasonably be attributed to $Y$. The coverage probability of $[x_R \pm 2u(x_R)]$ is at least 75% for any distribution with expected value $x_R$ and standard deviation $u(x_R)$ [8]. The coverage probability jumps to 95% for a normal distribution and 100% for a rectangular distribution. According to [1], 'Participation in a key comparison is open to laboratories having the highest technical competence and experience, normally the member laboratories of the appropriate CC'. Thus $x_1, \ldots, x_n$ are values attributed to $Y$ by competent laboratories. That is, $x_1, \ldots, x_n$ are plausible values for $Y$. Therefore, the interval $[x_R \pm 2u(x_R)]$ should exclude none or at most a small fraction of the results $x_1, \ldots, x_n$. When the number, $n$, of laboratories is large, the interval $[x_W \pm 2u(x_W)]$ may exclude[13] a large fraction of the results, $x_1, \ldots, x_n$. Therefore, the interval $[x_W \pm 2u(x_W)]$ may be unreasonably narrow for the uncertainty interval $[x_R \pm 2u(x_R)]$ associated with the key comparison reference value, $x_R$ [10].

The use of a simulated expected median, $x_M$, as the key comparison reference value, $x_R$, and $u(x_M)$ as the uncertainty, $u(x_R)$, is based on the measurement equation $Y = X_M$, where $X_M = m(X_1, \ldots, X_n)$ is the median function of the variables $X_1, \ldots, X_n$. This measurement equation does not include an input variable to represent a correction for possible bias in $x_M$. Therefore, the measurement equation $Y = X_M$ is incomplete. Consequently, the uncertainty, $u(x_M)$, is an underestimate. The bias in $x_M$ is the difference between the expected value of the sampling distribution of $x_M$ and the value, $Y$, of the measurand. The sampling distribution of $x_M$ describes the relative frequencies of occurrence for all possible values of $x_M$ if the CIPM key comparison could be repeated infinitely many times under exactly the same conditions. To the extent that the expected value of $x_M$ and the value $Y$ are unknown, one cannot be certain that the bias in $x_M$ is zero.

## 4. Systematic laboratory effects model

We will briefly describe a systematic laboratory effects model based on the ISO *Guide*. This model was proposed in [5] for the statistical analysis of a simple CIPM key comparison. In this model, the laboratory effects (biases), $X_i - Y$, for $i = 1, 2, \ldots, n$, are unknown constants. Consider a combined result of the form $\sum_i a_i x_i$, where $a_i \geqslant 0$ and $\sum_i a_i = 1$ that is used as a preliminary estimate[14] for the value $Y$ of the common measurand. We refer to the preliminary estimate as an uncorrected combined result (UCR) and denote it by $x_{UCR} = \sum_i a_i x_i$. If $a_i = w_i / \sum_i w_i$, then $x_{UCR}$ is the weighted mean $x_W = \sum_i w_i x_i / \sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. If $a_i = 1/n$ for $i = 1, 2, \ldots, n$, then $x_{UCR}$ is the arithmetic mean $x_A = \sum_i x_i / n$. Let $X_{UCR} = E(\sum_i a_i x_i) = \sum_i a_i E(x_i) = \sum_i a_i X_i$ be the expected value of the sampling distribution of $x_{UCR}$. The result, $x_{UCR}$, is subject to the bias $(X_{UCR} - Y)$. The ISO *Guide* recommends that the result, $x_{UCR}$, should be corrected to counter its possible bias and the uncertainty associated with the correction should be included in the combined standard uncertainty associated with the corrected result. The bias $(X_{UCR} - Y)$ is an unknown constant. The correction for bias, denoted by $C$, is a variable with a state-of-knowledge probability distribution[15]. Suppose the expected value and standard deviation of a state-of-knowledge probability distribution for the correction variable $C$ are $c$ and $u(c)$, respectively. Then the correction applied to the result $x_{UCR}$ to counter its possible bias is $c$ and the standard uncertainty associated with the correction is $u(c)$. In order to specify $c$ and $u(c)$, one is free to use any reasonable distribution for $C$ that has a finite expected value and a finite standard deviation. For example, one could use a rectangular distribution on the interval $(-\alpha, +\alpha)$ for some non-negative $\alpha$ specified by scientific judgment [4]. In that case $c = 0$ and $u(c) = \alpha/\sqrt{3}$. The uncertainty $u(c)$ can be made small or large by the choice of $\alpha$.

A measurement equation for $Y$ is required to incorporate a correction for possible bias in a result of measurement. The measurement equation that corresponds to the bias $(X_{UCR} - Y)$ in $x_{UCR}$ is $Y = X_{UCR} + C$. It suggests the following model for the value, $Y$, of the measurand:

$$\begin{aligned} E(X_i) &= x_i, & S(X_i) &= u(x_i), \\ X_{UCR} &= \sum_i a_i X_i, & Y &= X_{UCR} + C, \end{aligned} \tag{2}$$

where $a_1, \ldots, a_n$ are constants such that $a_i \geqslant 0$ and $\sum_i a_i = 1$. In this model, $X_1, \ldots, X_n, X_{UCR}, C$, and $Y$ are regarded as variables with state-of-knowledge distributions. The expected value and standard deviation[16] of $X_i$ are the given constants, $x_i$ and $u(x_i)$, respectively, for $i = 1, 2, \ldots, n$. A state-of-knowledge distribution for the correction variable $C$ is defined

[13] As $n$ increases, the expression $u(x_W) = 1/\sqrt{[\sum_i w_i]}$ decreases. Therefore, the fraction of results $x_1, \ldots, x_n$ excluded by the interval $[x_W \pm 2u(x_W)]$ increases. In the limit as $n$ tends to infinity, the interval $[x_W \pm 2u(x_W)]$ would exclude all the results $x_1, \ldots, x_n$.

[14] The results $x_1, \ldots, x_n$ are values attributed to $Y$ by competent laboratories. It is, therefore, not unreasonable to assume that the value $Y$ is either somewhere in the range of results $x_1, \ldots, x_n$ or in the vicinity of this range. The specifications $a_i \geqslant 0$ and $\sum_i a_i = 1$ associated with $\sum_i a_i x_i$ represent this assumption.

[15] A state-of-knowledge distribution for the correction $C$ implies a corresponding state-of-knowledge distribution for the bias $(X_{UCR} - Y)$. Since the input quantity that goes in the measurement equation is the correction variable, not the bias, we do not deal with a state-of-knowledge distribution for the bias $(X_{UCR} - Y)$.

[16] Ideally, the result, $x_i$, and uncertainty, $u(x_i)$, should be the expected value and standard deviation of a Bayesian posterior distribution for $X_i$. A Bayesian uncertainty has the advantage that it has no statistical uncertainty arising from a small number of measurements [8].

independent of the state-of-knowledge distributions for the variables $X_1, \dots, X_n$ after the expected values and standard deviations of the latter have been specified. Therefore, $X_i$ and $C$ are independently distributed, for $i = 1, 2, \dots, n$. Consequently, $X_{\mathrm{UCR}}$ and $C$ are independently distributed. The model (2) is referred to as a systematic laboratory effects model [5]. The systematic laboratory effects model (2) allows for the possibility that not all pairs of the variables $X_1, \dots, X_n$ may have independent state-of-knowledge distributions [5]. Suppose $r(x_i, x_j)$ is the *correlation coefficient* between $X_i$ and $X_j$ for $i, j = 1, \dots, n$ and $i \neq j$. Then the expected value and standard deviation of a state-of-knowledge distribution for $X_{\mathrm{UCR}}$ are, respectively, $E(X_{\mathrm{UCR}}) = \sum_i a_i E(X_i) = \sum_i a_i x_i = x_{\mathrm{UCR}}$ and $S(X_{\mathrm{UCR}}) = \surd[\sum_i a_i^2 u^2(x_i) + 2\sum_{(i<j)} a_i a_j u(x_i) u(x_j) r(x_i, x_j)] = u(x_{\mathrm{UCR}})$. The parameters $x_{\mathrm{UCR}}$ and $u(x_{\mathrm{UCR}})$ represent the centrality and spread of a state-of-knowledge distribution for the expected value, $X_{\mathrm{UCR}}$. The parameters $c$ and $u(c)$ represent the centrality and spread of a state-of-knowledge distribution for the correction $C$ for possible bias in $x_{\mathrm{UCR}}$. The CCR for $Y$ determined from the systematic laboratory effects model (2) is $y = x_{\mathrm{UCR}} + c$, and its associated standard uncertainty is $u(y) = \surd[u^2(x_{\mathrm{UCR}}) + u^2(c)]$. Thus the key comparison reference value, $x_{\mathrm{R}}$, based on model (2) is $y$ and $u(x_{\mathrm{R}})$ is $u(y)$.

Following the ISO *Guide*, we interpret $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ as the expected value and standard deviation of a state-of-knowledge distribution for $Y$ based on the data $x_1, \dots, x_n$ and $u(x_1), \dots, u(x_n)$. The coverage probability of the interval $[x_{\mathrm{R}} \pm 2u(x_{\mathrm{R}})]$ is the fraction of a state-of-knowledge distribution for $Y$ represented by $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ that is encompassed by this interval [4, 8].

The degrees of equivalence determined from the systematic laboratory effects model are $d_i = x_i - x_{\mathrm{R}} = x_i - y$ and $d_{i,j} = x_i - x_j$ for $i, j = 1, 2, \dots, n$ and $i \neq j$. The results, $x_1, \dots, x_n$, and $y$ are the expected values; and the uncertainties, $u(x_1), \dots, u(x_n)$ and $u(y)$ are the standard deviations of $X_1, \dots, X_n$ and $Y$. Therefore, the degree of equivalence, $d_i$, is the expected value of a state-of-knowledge distribution[17] for the laboratory effect (bias), $X_i - Y$, for $i = 1, 2, \dots, n$. The degree of equivalence, $d_{i,j}$, is the expected value of a state-of-knowledge distribution for $X_i - X_j$, the difference between the laboratory expected values $X_i$ and $X_j$, for $i, j = 1, 2, \dots, n$ and $i \neq j$. The uncertainty, $u(d_i)$, is the standard deviation[18] of $X_i - Y$, and the uncertainty, $u(d_{i,j})$, is the standard deviation of $X_i - X_j$, for $i, j = 1, 2, \dots, n$ and $i \neq j$. Thus the degrees of equivalence determined from the systematic laboratory effects model quantitate the agreements and disagreements between the laboratory results. Therefore, the systematic laboratory effects model is suitable for the data analysis of a simple CIPM key comparison.

*Note 1.* The measurement equation $Y = X_{\mathrm{UCR}} + C$ is defined by the chosen linear function $x_{\mathrm{UCR}} = \sum_i a_i x_i$, where $a_i \geqslant 0$ and $\sum_i a_i = 1$. Suppose the state-of-knowledge distributions for $X_1, \dots, X_n$ are independent. The systematic laboratory effects model is based on the assumption that the submitted uncertainties, $u(x_1), \dots, u(x_n)$, are all sufficiently reliable. It can be shown that the standard deviation $S(X_{\mathrm{UCR}}) = \surd[\sum_i a_i^2 u^2(x_i)]$ is minimum when $a_i = w_i / \sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \dots, n$. Therefore, one may prefer the measurement equation $Y = X_{\mathrm{W}} + C$, where $X_{\mathrm{W}} = \sum_i w_i X_i / \sum_i w_i$. When one is not certain that $u(x_1), \dots, u(x_n)$ are all sufficiently reliable, even thought this assumption is made, the measurement equation $Y = X_{\mathrm{A}} + C$, where $X_{\mathrm{A}} = \sum_i X_i / n$, may be preferred.

### *4.1. Probability distribution for the correction variable*

Reference [11] addressed the special case where $n = 2$ and $x_{\mathrm{UCR}} = x_{\mathrm{A}}$, which is in the middle of the two results $x_1$ and $x_2$. This work proposed for $C$ a rectangular distribution on the interval $(-\alpha, +\alpha)$, where $\alpha = |(x_1 - x_2)/2| = |x_{(1)} - x_{\mathrm{A}}| = |x_{(2)} - x_{\mathrm{A}}|$, $x_{(1)} = \min\{x_1, x_2\}$, and $x_{(2)} = \max\{x_1, x_2\}$; and a normal distribution with $c = 0$ and $u(c)$ such that $2u(c) = \alpha = |x_{(1)} - x_{\mathrm{A}}| = |x_{(2)} - x_{\mathrm{A}}|$. When $n$ is more than two, one may consider a rectangular distribution on the interval $(-\alpha, +\alpha)$, where $\alpha = \max\{|x_{(1)} - x_{\mathrm{UCR}}|, |x_{(n)} - x_{\mathrm{UCR}}|\}$, $x_{(1)} = \min\{x_1, \dots, x_n\}$, and $x_{(n)} = \max\{x_1, \dots, x_n\}$; and a normal distribution with $c = 0$ and $u(c)$ such that $2u(c) = \alpha = \max\{|x_{(1)} - x_{\mathrm{UCR}}|, |x_{(n)} - x_{\mathrm{UCR}}|\}$. When $n$ is more than two, $x_{\mathrm{UCR}}$ may not be in the middle of the results $x_1, \dots, x_n$. Therefore, the rectangular distribution with limits $\pm\alpha$ and the normal distribution with $c = 0$ and $2u(c) = \alpha$ may not fit the dispersion of results $x_1, \dots, x_n$. The uncertainty $u(c) = \alpha/\surd 3$ determined from a rectangular distribution on the interval $(-\alpha, +\alpha)$, where $\alpha = \max\{|x_{(1)} - x_{\mathrm{UCR}}|, |x_{(n)} - x_{\mathrm{UCR}}|\}$, may be too large. The bias in $x_{\mathrm{UCR}}$ is necessarily bounded; therefore, a normal distribution that is unbounded is an awkward probability distribution for $C$. Also, a normal distribution may give too much weight to the results near $x_{\mathrm{UCR}}$ and too little weight to the results far from $x_{\mathrm{UCR}}$. Therefore, in [10], we proposed for $C$ a rectangular distribution on the interval $(-\alpha_1, +\alpha_2)$, where $-\alpha_1 = (x_{(1)} - x_{\mathrm{UCR}})$ and $\alpha_2 = (x_{(2)} - x_{\mathrm{UCR}})$. The expected value and standard deviation of a rectangular distribution on the interval $(-\alpha_1, +\alpha_2)$ are $(\alpha_2 - \alpha_1)/2$ and $(\alpha_1 + \alpha_2)/\surd 12$, respectively. We also proposed, in [10], an asymmetric triangular distribution on the interval $(-\alpha_1, +\alpha_2)$, where $-\alpha_1 = (x_{(1)} - x_{\mathrm{UCR}})$, $\alpha_2 = (x_{(2)} - x_{\mathrm{UCR}})$, and the peak is at zero. The expected value and standard deviation of a triangular distribution on the interval $(-\alpha_1, +\alpha_2)$ with the peak at zero are $(\alpha_2 - \alpha_1)/3$ and $\surd[(\alpha_1 - \alpha_2)^2/18 + (\alpha_1\alpha_2)/6]$, respectively. Some of our colleagues criticized the asymmetric triangular distribution because it is determined by the extreme results, $x_{(1)}$ and $x_{(n)}$, which are sometimes suspected to be in error. So in [5], we proposed a discrete equal-probability distribution for $C$ that is determined by all of the results $x_1, \dots, x_n$. The correction $c$ and uncertainty $u(c)$ based on the discrete equal-probability distribution are $c = E(C) = x_{\mathrm{A}} - x_{\mathrm{UCR}}$ and $u(c) = S(C) = \surd[\sum_i (x_i - x_{\mathrm{A}})^2/n]$, respectively. When the discrete equal-probability distribution is used to specify $c$ and

[17] A state-of-knowledge distribution for the bias $X_i - Y$ requires a state-of-knowledge distribution for $Y$, which in turn requires a correction for bias in a UCR $x_{\mathrm{UCR}} = \sum_i a_i x_i$ for $Y$.

[18] The standard deviation of $X_i - Y$ depends on the covariance between $X_i$ and $Y$ for $i = 1, 2, \dots, n$. Since $Y = X_{\mathrm{UCR}} + C = \sum_i a_i X_i + C$ and the variable $C$ is distributed independent of the variables $X_1, \dots, X_n$, the covariances $C(X_i, Y)$, for $i = 1, 2, \dots, n$, can be determined from the variances and covariances of $X_1, \dots, X_n$. Then $u(d_i) = \surd[V(X_i - Y)]$, where the variance $V(X_i - Y)$ is equal to $V(X_i) + V(Y) - 2 \times C(X_i, Y)$. When the variables $X_1, \dots, X_n$ are independent, $C(X_i, Y) = a_i \times V(X_i)$ and $V(Y) = \sum_i a_i^2 V(X_i)$.

$u(c)$, then $y = x_{\mathrm{UCR}} + x_{\mathrm{A}} - x_{\mathrm{UCR}} = x_{\mathrm{A}}$. Thus the CCR, $y$, determined through the discrete equal-probability distribution is the arithmetic mean, $x_{\mathrm{A}}$, regardless of the linear function that is used as the UCR $x_{\mathrm{UCR}}$. When $x_{\mathrm{UCR}}$ is the arithmetic mean, $x_{\mathrm{A}}$, the correction $c$ is zero. In that case, $y = x_{\mathrm{UCR}} = x_{\mathrm{A}}$.

### *4.2. A mixture distribution for the value of the measurand*

Reference [12] suggests that the key comparison reference value, $x_{\mathrm{R}}$, and uncertainty, $u(x_{\mathrm{R}})$, may be determined from a mixture distribution for the value, $Y$, of the measurand based on the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. The result, $x_i$, and uncertainty, $u(x_i)$, are interpreted as the expected value and standard deviation of a probability distribution attributed to $Y$ in the laboratory labelled $i$ for $i = 1, 2, \ldots, n$. The key comparison reference value, $x_{\mathrm{R}}$, and uncertainty, $u(x_{\mathrm{R}})$, are the expected value and standard deviation of a combined probability distribution attributed to $Y$. Suppose the probability density function (PDF) represented by $x_i$ and $u(x_i)$ is $p_i(y)$ and the PDF represented by $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ is $p(y)$. The PDF $p(y)$ may be defined as a linear combination $p(y) = \sum_i \kappa_i p_i(y)$ of the PDFs $p_i(y)$, where $\kappa_1, \ldots, \kappa_n$ are non-negative weights attributed to $p_1(y), \ldots, p_n(y)$ such that $\sum_i \kappa_i = 1$. The combined probability distribution with PDF $p(y) = \sum_i \kappa_i p_i(y)$ is referred to as a mixture distribution. The expected value and standard deviation of $p(y)$ are $\sum_i \kappa_i x_i$ and $\surd[\sum_i \kappa_i u^2(x_i) + \sum_i \kappa_i (x_i - \sum_i \kappa_i x_i)^2]$ [13]. Reference [12] suggests that $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ may be defined by setting $\kappa_i = 1/n$ for $i = 1, 2, \ldots, n$. Then $x_{\mathrm{R}} = \sum_i x_i/n = x_{\mathrm{A}}$, $u(x_{\mathrm{R}}) = \surd[(1/n)\sum_i u^2(x_i) + \sum_i (x_i - x_{\mathrm{A}})^2/n]$, and the interval $[x_{\mathrm{R}} \pm 2u(x_{\mathrm{R}})]$ represents an approximate range of the values that could reasonably be attributed to $Y$ based on the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. The corresponding expressions[19] for $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ determined from the systematic laboratory effects model are $x_{\mathrm{R}} = x_{\mathrm{A}}$ and $u(x_{\mathrm{R}}) = \surd[(1/n^2)\sum_i u^2(x_i) + \sum_i (x_i - x_{\mathrm{A}})^2/n]$. The only difference between these two expressions for $u(x_{\mathrm{R}})$ is the coefficient of $\sum_i u^2(x_i)$, which is $1/n$ from the mixture distribution and $1/n^2$ from the systematic laboratory effects model[20]. The systematic laboratory effects model is more flexible than a mixture distribution because the UCR may be any linear combination $x_{\mathrm{UCR}} = \sum_i a_i x_i$ of the results $x_1, \ldots, x_n$, where $a_i \geqslant 0$ and $\sum_i a_i = 1$, and the correction variable $C$ may have any reasonable probability distribution. In addition, the systematic laboratory effects model allows for the possibility that not all pairs of the variables $X_1, \ldots, X_n$ may have independent state-of-knowledge distributions.

### *4.3. Standardized degrees of equivalence*

Generally, the submitted uncertainties $u(x_1), \ldots, u(x_n)$ are unequal; therefore, the uncertainties $u(d_1), \ldots, u(d_n)$ are different. When the uncertainties $u(d_1), \ldots, u(d_n)$ are different, one may compare the dimensionless ratios $d_1/u(d_1), \ldots, d_n/u(d_n)$. The dimensionless ratio for a laboratory that submits a too small uncertainty would be inflated, and a laboratory that submits a too large uncertainty would be deflated. Therefore, we suggest that the degrees of equivalence, $d_1, \ldots, d_n$, be divided by a common uncertainty. In particular, we suggest that the degrees of equivalence, $d_1, \ldots, d_n$, be divided by the uncertainty, $u(x_{\mathrm{R}})$. Therefore, we propose the dimensionless expression $E_i = (x_i - x_{\mathrm{R}})/u(x_{\mathrm{R}})$ as the *standardized degree of equivalence* of the result $x_i$ for $i = 1, 2, \ldots, n$. The corresponding expression for the *standardized degree of equivalence* of $x_i$ and $x_j$ is $E_i - E_j = [(x_i - x_{\mathrm{R}})/u(x_{\mathrm{R}}) - (x_j - x_{\mathrm{R}})/u(x_{\mathrm{R}})]$ for $i, j = 1, 2, \ldots, n$ and $i \neq j$ [10].

## 5. Statistical analysis based on the systematic laboratory effects model

We assume that the submitted results $x_1, \ldots, x_n$ are mutually comparable and appropriate for determining the key comparison reference value, and the submitted uncertainties $u(x_1), \ldots, u(x_n)$ are sufficiently reliable. It is useful to classify CIPM key comparisons according to the type of results, $x_1, \ldots, x_n$, that are compared. *Comparison of the first kind*: the laboratory results, $x_1, \ldots, x_n$, are direct measurements of a common measurand of a stable value, $Y$, during the comparison. *Comparison of the second kind*: the laboratory results, $x_1, \ldots, x_n$, are not direct measurements of a stable measurand. Many CIPM key comparisons are of the second kind because it is often difficult or impossible to realize exactly the same measurand for all participants.

### *5.1. Statistical analysis for a comparison of the first kind*

The key comparison reference value, $x_{\mathrm{R}}$, and uncertainty, $u(x_{\mathrm{R}})$, are identified with the CCR, $y$, and uncertainty, $u(y)$, respectively. The following three steps are required to determine $y$ and $u(y)$ from the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ through the systematic laboratory effects model. First, determine a UCR, $x_{\mathrm{UCR}}$, and its associated standard uncertainty, $u(x_{\mathrm{UCR}})$. Second, determine the correction $c$ and uncertainty $u(c)$ to counter possible bias $(X_{\mathrm{UCR}} - Y)$ in $x_{\mathrm{UCR}}$. Third, determine the CCR, $y$, and its associated combined standard uncertainty, $u(y)$.

*Note 1.* When a number of travelling standards are circulated, the value $Y$ may be defined as the average value of the travelling standards. In that case, the results, uncertainties, and correlation coefficients are determined from the measurements for all travelling standards.

*Note 2.* In some CIPM key comparisons the values of the travelling standards may drift in a recognized time dependent manner. Such data may be analysed as a CIPM key comparison of the first kind. The pilot laboratory periodically measures the travelling standards during the entire duration of the comparison to quantify their drift. The participants measure the travelling standards at different time periods. The measured results are adjusted for the drift using the measurements made by the pilot laboratory to determine the laboratory results $x_1, \ldots, x_n$ [14]. The uncertainties $u(x_1), \ldots, u(x_n)$ must include the components of uncertainty associated with the adjustments. The uncertainty, $u(x_{\mathrm{UCR}})$, associated with the

[19] The corresponding UCR, $x_{\mathrm{UCR}}$, is the arithmetic mean, $x_{\mathrm{A}}$, and the probability distribution for the correction variable $C$ is the discrete equal-probability distribution.

[20] Therefore, the uncertainty, $u(x_{\mathrm{R}})$, determined from the systematic laboratory effects model is less than the uncertainty, $u(x_{\mathrm{R}})$, determined from the mixture distribution.

**Table 1.** Relative differences, $x_i$, between individual laboratory and BIPM measurements and standard uncertainties, $u(x_i)$, for short ($S$), medium ($M$), and long ($L$) wavelengths.

| Laboratory indices and names | Wavelength $S$ | | Wavelength $M$ | | Wavelength $L$ | |
|---|---|---|---|---|---|---|
| | $x_i \times 10^4$ | $u(x_i) \times 10^4$ | $x_i \times 10^4$ | $u(x_i) \times 10^4$ | $x_i \times 10^4$ | $u(x_i) \times 10^4$ |
| 1. ptb.t | −0.80 | 1.3 | −0.2 | 1.3 | 0.0 | 1.3 |
| 2. bnm.inm | 1.80 | 2.0 | 1.1 | 1.7 | 0.6 | 1.4 |
| 3. csiro | 1.50 | 1.4 | 2.0 | 1.4 | 1.35 | 1.4 |
| 4. dfm | −0.45 | 2.5 | −0.3 | 2.5 | −0.5 | 2.5 |
| 5. etl | 15.10 | 4.9 | 13.1 | 4.9 | 17.15 | 4.9 |
| 6. hut | 2.30 | 2.7 | 1.7 | 2.7 | −0.4 | 2.7 |
| 7. ien | −17.60 | 6.8 | −11.0 | 6.8 | 0.7 | 6.8 |
| 8. ifa | 3.30 | 2.2 | 0.0 | 2.2 | −1.3 | 2.2 |
| 9. msl | 0.40 | 1.2 | 0.3 | 1.3 | 0.6 | 1.4 |
| 10. kriss | −1.25 | 2.4 | −5.1 | 2.4 | −0.65 | 2.4 |
| 11. nist | 7.30 | 4.5 | 5.9 | 3.2 | 2.9 | 4.2 |
| 12. nmi.vsl | −1.45 | 2.6 | −1.1 | 2.6 | −0.55 | 2.6 |
| 13. npl | −0.30 | 1.1 | 1.3 | 1.1 | 1.4 | 1.2 |
| 14. nrc | 3.00 | 3.4 | 5.3 | 3.4 | 4.6 | 3.4 |
| 15. ptb.r | 3.20 | 2.1 | 2.9 | 2.9 | 2.5 | 1.4 |
| 16. sp | −1.10 | 5.1 | −1.0 | 5.1 | −1.3 | 5.1 |

UCR, $x_{\mathrm{UCR}}$, must include the correlation coefficients $r(x_i, x_j)$, for $i, j = 1, \ldots, n$ and $i \neq j$.

*Note 3.* When the measurements are made at various settings of a parameter such as wavelength or frequency, the calculations are repeated for each wavelength or frequency.

### 5.2. *Statistical analysis for a comparison of the second kind*

The results $x_1, \ldots, x_n$ are not direct measurements of a stable measurand of value $Y$. We will define the value $Y$ for a CIPM key comparison of the second kind and then interpret $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$ in terms of that value. The subject experts judiciously design a CIPM key comparison such that the results $x_1, \ldots, x_n$ are mutually comparable. Therefore, we may define the value $Y$ for a CIPM key comparison of the second kind as a *statistical prediction similar to the results $x_1, \ldots, x_n$ that might be realized by a competent laboratory similar to the laboratories that participated in the comparison.* With this interpretation of the value $Y$, the measurement equation $Y = X_{\mathrm{UCR}} + C$ may be used for a CIPM key comparison of the second kind. Thus the CCR, $y$, for the prediction $Y$ and uncertainty $u(y)$ may be determined from the systematic laboratory effects model. The key comparison reference value, $x_{\mathrm{R}}$, and uncertainty, $u(x_{\mathrm{R}})$, are identified with $y$ and $u(y)$, respectively, and interpreted as the expected value and standard deviation of a state-of-knowledge distribution for the values that could reasonably be attributed to the prediction $Y$. In particular, the key comparison reference value, $x_{\mathrm{R}}$, is a value that could reasonably be attributed to the prediction $Y$ of the result that might be realized by a competent laboratory. In section 7, we illustrate the statistical analysis for a comparison of the second kind.

## 6. A useful method for identifying discrepant results and uncertainties

According to the Statistical Guidelines paper, the data published in the official Draft A of the CIPM key comparison should be used to determine the key comparison reference value, $x_{\mathrm{R}}$, and uncertainty, $u(x_{\mathrm{R}})$. However, some CCs may choose to screen and possibly adjust the data published in the Draft A before using them to determine $x_{\mathrm{R}}$ and $u(x_{\mathrm{R}})$. Even though the participants of a CIPM key comparison are competent laboratories, the uncertainties stated by some may be unreasonably small in view of the experts. A practical remedy concerning unreasonably small uncertainties is to replace them with a more reasonable 'cut-off' value determined by the experts. Also, one or more of the results may seem erroneous in view of the previous or other measurements. A decision concerning seemingly erroneous results is a matter of expert judgment and is subject to the protocol and the limitations of time and resources available for investigation.

**Table 2.** The $k$-statistic and the $h$-statistic for the data shown in table 1.

| Laboratory indices and names | Wavelength $S$ | | Wavelength $M$ | | Wavelength $L$ | |
|---|---|---|---|---|---|---|
| | $k$ | $h$ | $k$ | $h$ | $k$ | $h$ |
| 1. ptb.t | 0.395 | −0.269 | 0.403 | −0.222 | 0.402 | −0.383 |
| 2. bnm.inm | 0.607 | 0.134 | 0.526 | 0.033 | 0.433 | −0.247 |
| 3. csiro | 0.425 | 0.088 | 0.434 | 0.210 | 0.433 | −0.078 |
| 4. dfm | 0.759 | −0.215 | 0.774 | −0.242 | 0.773 | −0.496 |
| 5. etl | 1.487 | 2.196 | 1.518 | 2.392 | 1.515 | 3.494 |
| 6. hut | 0.819 | 0.212 | 0.836 | 0.151 | 0.835 | −0.473 |
| 7. ien | 2.064 | −2.874 | 2.106 | −2.345 | 2.103 | −0.225 |
| 8. ifa | 0.668 | 0.367 | 0.681 | −0.183 | 0.680 | −0.677 |
| 9. msl | 0.364 | −0.083 | 0.403 | −0.124 | 0.433 | −0.247 |
| 10. kriss | 0.728 | −0.339 | 0.743 | −1.185 | 0.742 | −0.530 |
| 11. nist | 1.366 | 0.987 | 0.991 | 0.977 | 1.299 | 0.273 |
| 12. nmi.vsl | 0.789 | −0.370 | 0.805 | −0.399 | 0.804 | −0.507 |
| 13. npl | 0.334 | −0.191 | 0.341 | 0.072 | 0.371 | −0.066 |
| 14. nrc | 1.032 | 0.320 | 1.053 | 0.859 | 1.051 | 0.657 |
| 15. ptb.r | 0.637 | 0.351 | 0.898 | 0.387 | 0.433 | 0.182 |
| 16. sp | 1.548 | −0.315 | 1.579 | −0.380 | 1.577 | −0.677 |

A useful statistical method for flagging discrepant results $x_1, \ldots, x_n$ and uncertainties $u(x_1), \ldots, u(x_n)$ is the ASTM documentary standard E691-1999[21] [15] or its equivalent.

[21] Author names are not associated with the ASTM standards. However, the original issue of the ASTM standard E691, dated 1979, was drafted by Dr John Mandel and Dr Robert Paule based on their experience with interlaboratory

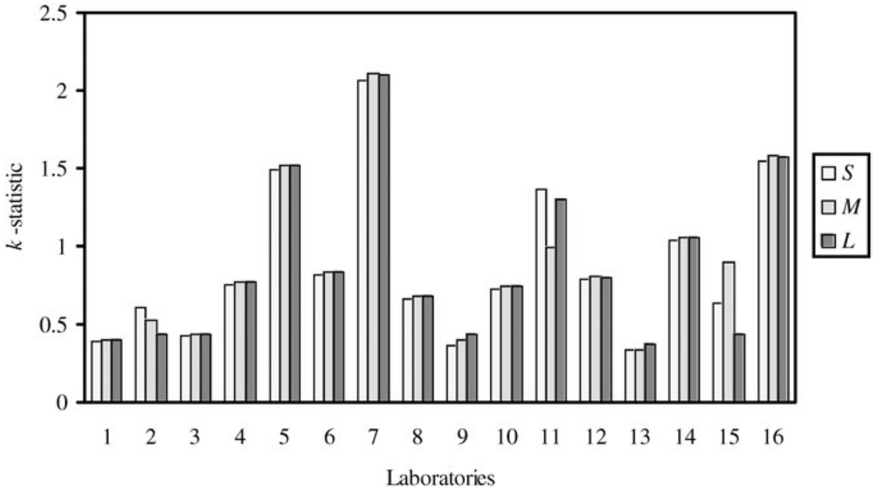

**Figure 1.** Chart of the $k$-statistic for the data shown in table 1.

The ASTM standard E691 requires that all $n$ laboratories make measurements on $m$ different materials. The ASTM standard E691 refers to the set of measurements from a particular laboratory and for a particular material as a cell. There are $n \times m$ cells. A $k$-statistic and an $h$-statistic are calculated for each cell. The $k$-statistic is defined as $\sqrt{[u^2(x_i)/(\sum_i u^2(x_i)/n)]}$; i.e. the $k$-statistic is the *normalized cell uncertainty*. It is used to compare the uncertainties $u(x_1), \ldots, u(x_n)$. The $h$-statistic is defined as $(x_i - x_A)/s$, where $s = \sqrt{[\sum_i (x_i - x_A)^2/(n-1)]}$; i.e. the $h$-statistic is the *standardized cell arithmetic mean*. It is used to compare the results $x_1, \ldots, x_n$. Charts of the $k$-statistic and the $h$-statistic display the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ in ways that make it easy to check for discrepant uncertainties and discrepant results. The ASTM standard E691 describes statistical tests of discrepancy for measurements that have approximately normal sampling distributions.

### *6.1. An example illustrating the use of the ASTM standard E691*

For illustration, we have used the data from the supplementary comparison of cryogenic radiometers CCPR S3 [16] that we had previously used in [10]. We may think of the wavelengths as different materials. Since not all laboratories used all six wavelengths, the ASTM standard E691 cannot directly be used for the CCPR S3 data. So we have slightly modified the data as discussed below. The modified data are displayed in table 1 for $n = 16$ laboratories and $m = 3$ wavelengths. The results $x_1, \ldots, x_{16}$ for the first 16 laboratories in [16] are the relative differences from the BIPM measurement and the last result, $x_{17}$, from the BIPM is identically equal to zero. So the two subsets $\{x_1, \ldots, x_{16}\}$ and $\{x_{17}\}$ of the results $x_1, \ldots, x_{17}$ represent different quantities. We have excluded the BIPM result, $x_{17}$, reducing the number, $n$, of laboratories to 16. Each laboratory used one or both of the wavelengths 476 nm and 488 nm. Each laboratory used the wavelength 514 nm. Each laboratory used one or both of the wavelengths 633 nm and 647 nm. The modified data are for three wavelengths, $S$ (short), $M$ (medium), and $L$ (long). The data for the short wavelength, $S$, are merged results for the wavelengths 476 nm and 488 nm. For those laboratories that used both wavelengths, the data are the average of the two results; for those laboratories that used only one of the two wavelengths, the data are results for the wavelength that was used. The data for the long wavelength, $L$, are similarly merged results for the wavelengths 633 nm and 647 nm. The data for the medium wavelength, $M$, are the results for the wavelength 514 nm, which was used by all laboratories. The data for the wavelength 568 nm are not used. The uncertainties given in table 1 are reproduced from table 65 of [16].

For the data in table 1, the computed values of the $k$-statistic and the $h$-statistic are given in table 2. The charts of the $k$-statistic and the $h$-statistic are shown in figures 1 and 2, respectively. From figure 1, we note that the uncertainties from laboratories 5, 7, and 16 are much larger and the uncertainties from laboratories 11 and 14 are somewhat larger than the rest. From figure 2, we note that the results for laboratories 5 and 7 seem to be different from the rest. The laboratories 5 and 7 are implicated in both figures 1 and 2.

The data for the wavelength 514 nm for 14 laboratories, excluding laboratories 5 and 7, are reproduced in table 3. These are original data from [16]. In [10], we had used the entire data for all 17 laboratories. Now we are using a subset of the data.

## 7. Statistical analysis of the data from the supplementary comparison CCPR S3

We have used the data from table 3 to calculate the UCR, $x_{\text{UCR}}$, correction, $c$, CCR, $y$, and their associated uncertainties using the triangular and the discrete equal-probability distributions for the correction variable $C$. We have chosen the arithmetic mean, $x_A$, as the UCR, $x_{\text{UCR}}$. Another choice is the weighted mean, $x_W$. In CCPR S3, the result $x_i$ is the average relative difference in the responsivity for a set of transfer standard detectors calibrated at the laboratory labelled $i$ and the same

evaluations at the predecessor organizations of the Chemical Science and Technology Laboratory (CSTL) of the National Institute of Standards and Technology (NIST), US Department of Commerce. The original objectives of the ASTM standard E691 are assessment of a test method and quantification of its repeatability and reproducibility standard deviations. We suggest its use for flagging discrepant results and uncertainties.

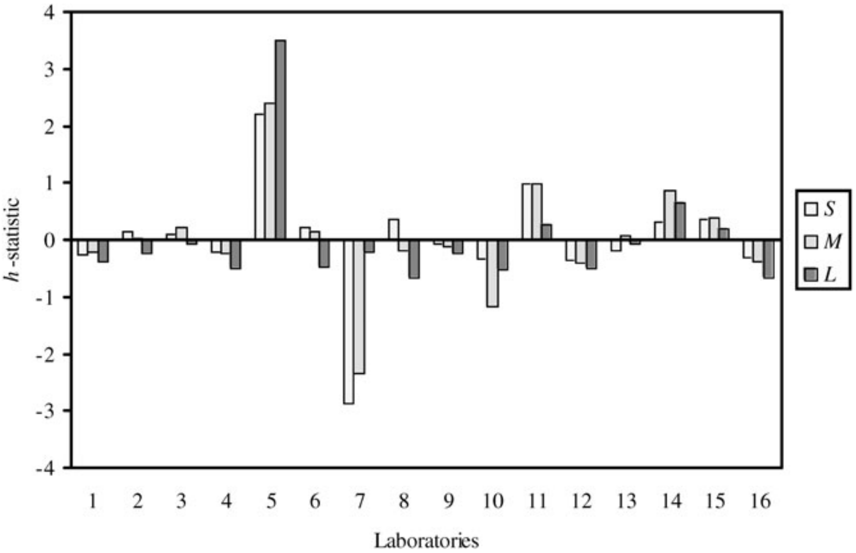

**Figure 2.** Chart of the $h$-statistic for the data shown in table 1.

**Table 3.** Relative differences, $x_i$, between individual laboratory and BIPM measurements and standard uncertainties, $u(x_i)$, for the wavelength 514 nm excluding the laboratories labelled 5 and 7.

| Laboratory indices and names | Wavelength 514 nm | |
|---|---|---|
| | $x_i \times 10^4$ | $u(x_i) \times 10^4$ |
| 1. ptb.t | −0.2 | 1.3 |
| 2. bnm.inm | 1.1 | 1.7 |
| 3. csiro | 2.0 | 1.4 |
| 4. dfm | −0.3 | 2.5 |
| 6. hut | 1.7 | 2.7 |
| 8. ifa | 0.0 | 2.2 |
| 9. msl | 0.3 | 1.3 |
| 10. kriss | −5.1 | 2.4 |
| 11. nist | 5.9 | 3.2 |
| 12. nmi.vsl | −1.1 | 2.6 |
| 13. npl | 1.3 | 1.1 |
| 14. nrc | 5.3 | 3.4 |
| 15. ptb.r | 2.9 | 2.9 |
| 16. sp | −1.0 | 5.1 |

**Table 4.** The arithmetic mean, $x_A$, and uncertainty, $u(x_A)$, the correction, $c$, and uncertainty, $u(c)$, and the CCR, $y$, and uncertainty, $u(y)$, determined from triangular and discrete equal-probability distributions for the data shown in table 3.

| Component | Result | Standard uncertainty |
|---|---|---|
| Uncorrected combined result | $x_A \times 10^4 = 0.91$ | $u(x_A) \times 10^4 = 0.70$ |
| Correction based on triangular distribution | $c \times 10^4 = -0.34$ | $u(c) \times 10^4 = 2.25$ |
| Corrected combined result from triangular distribution | $y \times 10^4 = 0.57$ | $u(y) \times 10^4 = 2.36$ |
| Correction based on discrete equal-probability distribution | $c \times 10^4 = 0.00$ | $u(c) \times 10^4 = 2.64$ |
| Corrected combined result from discrete equal-probability distribution | $y \times 10^4 = 0.91$ | $u(y) \times 10^4 = 2.74$ |

detectors calibrated at the BIPM, for $i = 1, 2, \ldots, n$. Thus CCPR S3 is a comparison of the second kind. The value $Y$ corresponding to the results $x_1, \ldots, x_n$ may be defined as a statistical prediction similar to the results $x_1, \ldots, x_n$ that might be realized by a competent laboratory similar to the laboratories that participated in the supplementary comparison CCPR S3. With this interpretation of the value $Y$ corresponding to the results $x_1, \ldots, x_n$, the systematic laboratory effects model may be applied to the data shown in table 3. Thus the result, $y$, and uncertainty, $u(y)$, may be determined from the systematic laboratory effects model. Table 4 shows $x_A$, $u(x_A)$, $c$, $u(c)$, $y$, and $u(y)$ determined from the triangular distribution with default limits $(x_{(1)} - x_A)$ and $(x_{(n)} - x_A)$ and the peak at zero, and the discrete equal-probability distribution for the correction variable $C$.

Figures 3, 4, and 5 display the 14 results, $x_1, \ldots, x_n$, from table 3 and their two-standard uncertainty intervals $[x_i \pm 2u(x_i)]$ for $i = 1, 2, \ldots, n$. Figure 3 plots the uncertainty interval $[x_A \pm 2u(x_A)]$, where $u(x_A)$ is the uncertainty associated with the UCR, $x_A$. Figure 4 plots the uncertainty interval $[y \pm 2u(y)]$ determined from the triangular distribution. Figure 5 plots the uncertainty interval $[y \pm 2u(y)]$ determined from the discrete equal-probability distribution. Figures 4 and 5 are similar but different from figure 3. The intervals $[y \pm 2u(y)]$ shown in figures 4 and 5 represent the CCPR S3 reference value and its associated uncertainty determined from the systematic laboratory effects model. The statistical interpretation of these intervals is consistent with the ISO *Guide*.

## 8. Summary

An international advisory group commissioned by the director of the BIPM has published guidelines for the statistical analysis of a simple CIPM key comparison where a travelling standard of a stable value during the comparison

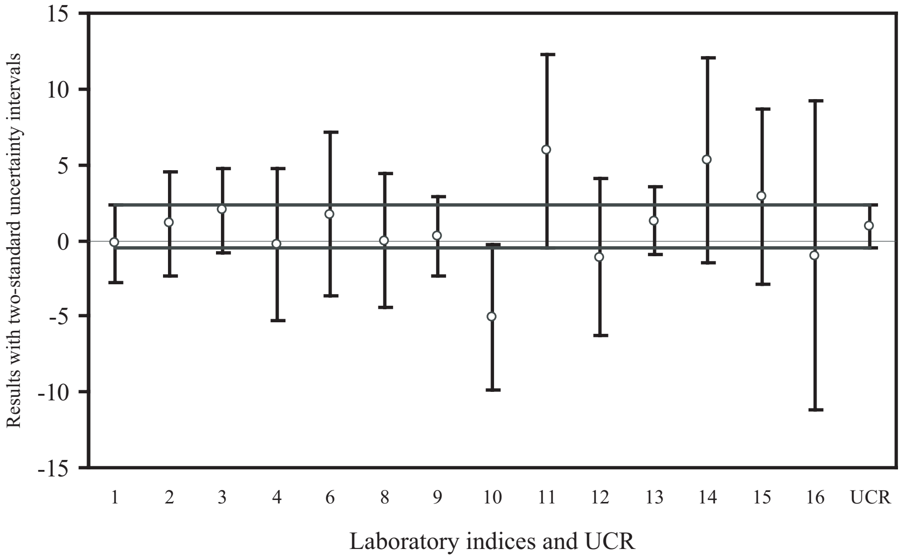

**Figure 3.** Results from table 3 with their arithmetic mean as UCR shown on the right.

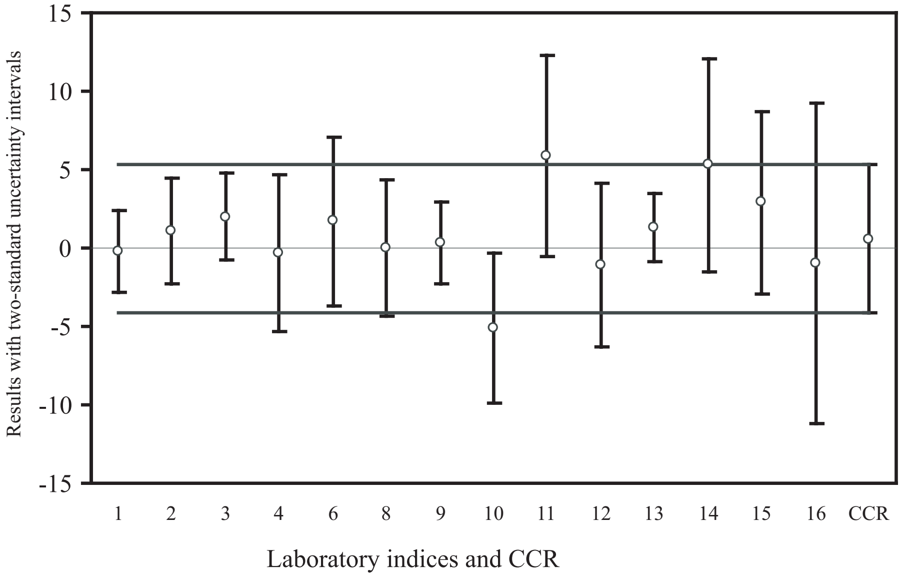

**Figure 4.** Results from table 3 with CCR, determined using the triangular distribution with limits $(x_{(1)} - x_A)$ and $(x_{(n)} - x_A)$, shown on the right.

is independently measured by the participants. There are a number of concerns regarding these guidelines. Also, many CIPM key comparisons are not simple. So the guidelines do not apply to many comparisons. This paper has introduced a systematic laboratory effects model for the statistical analysis of CIPM key comparisons. This model applies to all those comparisons where the data are appropriate for determining the key comparison reference value.

The data submitted by the participating laboratories consist of the paired results and standard uncertainties $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. The value $Y$ is either the value of a stable measurand or a statistical prediction similar to the results, $x_1, \ldots, x_n$, that might be realized by a competent laboratory similar to the laboratories that participated in the comparison. The ISO *Guide*'s eight steps ([4], section 8) for determining $x_R$, $u(x_R)$, and an uncertainty interval $[x_R \pm ku(x_R)]$ for $Y$ based on the systematic laboratory effects model are as follows.

*Step 1*. The measurement equation for $Y$ corresponding to an additive bias is $Y = X_{UCR} + C$, where $X_{UCR} = \sum_i a_i X_i$, $a_i \geqslant 0$, $\sum_i a_i = 1$, $X_1, \ldots, X_n$ are the laboratory expected values, and $C$ is a correction for possible bias in the UCR $x_{UCR} = \sum_i a_i x_i$. Here, $X_1, \ldots, X_n$, $X_{UCR}$, $C$, and $Y$ are variables with state-of-knowledge distributions. The alternatives for $x_{UCR}$ include the weighted mean, $x_W$, and the arithmetic mean, $x_A$, of the results $x_1, \ldots, x_n$. A distribution for $C$ is defined independent of the distributions for $X_1, \ldots, X_n$.

*Step 2*. The results, $x_1, \ldots, x_n$, are regarded as the expected values of the state-of-knowledge distributions for $X_1, \ldots, X_n$, respectively. Specify a reasonable distribution for $C$ based on scientific judgment. The alternatives for the distribution

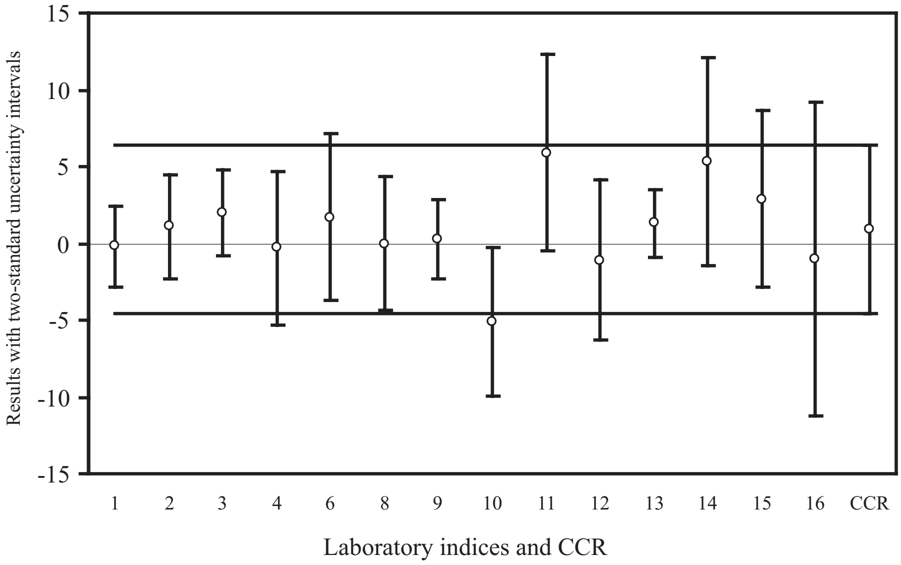

**Figure 5.** Results from table 3 with CCR, determined using the discrete equal-probability distribution, shown on the right.

of $C$ include a rectangular distribution on some interval [4], asymmetric triangular distribution [10], and discrete equal-probability distribution [5]. The expected value, $E(C)$, of the distribution for $C$ is denoted by $c$.

*Step 3.* The uncertainties, $u(x_1), \ldots, u(x_n)$, are regarded as the standard deviations of the state-of-knowledge distributions for $X_1, \ldots, X_n$, respectively. The standard deviation, $S(C)$, of the distribution for $C$ is denoted by $u(c)$.

*Step 4.* Quantify the correlation coefficients, $r(x_i, x_j)$, for the pairs $X_i$ and $X_j$ that might be correlated, where $i, j = 1, \ldots, n$ and $i \neq j$.

*Step 5.* The CCR for $Y$ is $y = x_{\mathrm{UCR}} + c$. The result, $y$, is identified with the key comparison reference value, $x_{\mathrm{R}}$.

*Step 6.* The standard uncertainty associated with $y$ is $u(y) = \sqrt{[}u^2(x_{\mathrm{UCR}}) + u^2(c)]$, where $u(x_{\mathrm{UCR}}) = \sqrt{[}\sum_i a_i^2 u^2(x_i) + 2\sum_{(i<j)} a_i a_j u(x_i) u(x_j) r(x_i, x_j)]$. The uncertainty, $u(y)$, is identified with the standard uncertainty, $u(x_{\mathrm{R}})$.

*Step 7.* If it is necessary to express the uncertainty as an interval, multiply the combined standard uncertainty $u(y) \equiv u(x_{\mathrm{R}})$ by a coverage factor, $k$, to obtain the interval $[y \pm ku(y)] \equiv [x_{\mathrm{R}} \pm ku(x_{\mathrm{R}})]$. The conventional value of $k$ is two.

*Step 8.* Report the key comparison reference value, $x_{\mathrm{R}}$, the standard uncertainty, $u(x_{\mathrm{R}})$, and the uncertainty interval $[x_{\mathrm{R}} \pm ku(x_{\mathrm{R}})]$.

The key comparison reference value, $x_{\mathrm{R}}$, and its associated standard uncertainty, $u(x_{\mathrm{R}})$, are interpreted as the expected value and standard deviation of a state-of-knowledge distribution for the values that could reasonably be attributed to $Y$ based on the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. The degree of equivalence $d_i = x_i - x_{\mathrm{R}} = x_i - y$ is the expected value of a state-of-knowledge distribution for the laboratory effect (bias) $X_i - Y$, and the uncertainty, $u(d_i)$, is the standard deviation of $X_i - Y$, for $i = 1, 2, \ldots, n$. The degree of equivalence $d_{i,j} = x_i - x_j$ is the expected value of a state-of-knowledge distribution for $X_i - X_j$, the difference between the laboratory expected values, $X_i$ and $X_j$, and the uncertainty $u(d_{i,j})$ is the standard deviation of $X_i - X_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$.

## Acknowledgments

The views expressed in this paper are of the authors and not necessarily the consensus of the institute they represent. The following provided helpful comments on earlier drafts of this paper: Ted Vorburger, Ron Boisvert, Eric Shirley, Tony Kearsley, Jim Gardener, Nell Sedransk, and Al Jones.

## References

[1] *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes* 1999 International Committee of Weights and Measures (CIPM) http://www1.bipm.org/utils/en/pdf/mra_2003.pdf

[2] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589–95

[3] *Guidelines for CIPM Key Comparisons* 1999 International Committee of Weights and Measures (CIPM) http://www1.bipm.org/utils/en/pdf/guidelines.pdf

[4] *Guide to the Expression of Uncertainty in Measurement* 1995 2nd edn (Geneva: International Organization for Standardization) ISBN 92-67-10188-9

[5] Kacker R N, Datla R U and Parr A C 2003 Statistical interpretation of key comparison reference value and degrees of equivalence *J. Res. Natl Inst. Stand. Technol.* **108** 439–46

[6] Kackar R N and Harville D A 1984 Approximations for standard errors of estimators of fixed and random effects in mixed linear models *J. Am. Stat. Assoc.* **79** 853–62

[7] Searle S R, Casella G and McCulloch C E 1992 *Variance Components* (New York: Wiley)

[8] Kacker R N and Jones A T 2003 On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent *Metrologia* **40** 235–48

[9] BIPM key comparison data base http://kcdb.bipm.org

[10] Kacker R N, Datla R U and Parr A C 2002 Combined result and associated uncertainty from interlaboratory evaluations based on the ISO Guide *Metrologia* **39** 279–93

[11] Levenson M S, Banks D L, Eberhardt K R, Gill L M, Guthrie W F, Liu H K, Vangel M G, Yen J H and Zhang N F 2000 An approach to combining results from multiple methods motivated by the ISO GUM *J. Res. Natl Inst. Stand. Technol.* **105** 571–9

[12] Toman B 2004 A Bayesian approach to assessing uncertainty and calculating a reference value in key comparison experiments, submitted, http://www.itl.nist.gov/div898/bios/toman.html

[13] Stuart A and Ord J K 1987 *Kendall's Advanced Theory of Statistics: Distribution Theory* 5th edn (New York: Oxford University Press)

[14] Jeffery A M 2002 Final report on key comparison CCEM-K4 of 10 pF capacitance standards *Metrologia* **39** (Tech. Suppl.) 01003 http://www.bipm.org/utils/common/pdf/final_reports/EM/K4/CCEM-K4.pdf

[15] *Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method* 1999 ASTM standard E691 http://www.astm.org

[16] Goebel R, Stock M and Köhler R 2000 Report on the international comparison of cryogenic radiometers based on transfer detectors *BIPM Rapport-2000/9* (Sèvres: Bureau International des Poids et Mesures)