[J. Res. Natl. Inst. Stand. Technol. **116**, 809-820 (2011)]

Combining Results From Multiple Evaluations of the Same Measurand

Volume 116 Number 6 November-December 2011

Rüdiger Kessel and Raghu N. Kacker

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

and

Klaus-Dieter Sommer

Physikalisch-Technische Bundesanstalt,
Braunschweig D-38116
Germany

ruediger.kessel@nist.gov
raghu.kacker@nist.gov
klaus-dieter.sommer@ptb.de

According to the Guide to the Expression of Uncertainty in Measurement (GUM), a result of measurement consists of a measured value together with its associated standard uncertainty. The measured value and the standard uncertainty are interpreted as the expected value and the standard deviation of a state-of-knowledge probability distribution attributed to the measurand. We discuss the term metrological compatibility introduced by the International Vocabulary of Metrology, third edition (VIM3) for lack of significant differences between two or more results of measurement for the same measurand. Sometimes a combined result of measurement from multiple evaluations of the same measurand is needed. We propose an approach for determining a combined result which is metrologically compatible with the contributing results.

Key words: consensus value; inter-laboratory evaluations; uncertainty in measurement.

Accepted: July 20, 2011

Available online: http://www.nist.gov/jres

1. Introduction

A function of various calibration laboratories, measurement standards organizations, national metrology institutes (NMIs), and international organizations such as the International Bureau of Weights and Measures (BIPM), the International Organization for Standardization (ISO), the International Organization of Legal Metrology (OIML), and the International Electro-technical Commission (IEC) is to ensure that the differences are insignificant between different measured values for the same measurand determined in various places, at various times, and by various measurement procedures. Without this assurance, the world's commerce, trade, manufacturing, engineering, and scientific research would be chaotic.

The old-time thinking concerning the uncertainty in measurement based on statistical error analysis are inappropriate for the rapidly advancing science and technology of measurement. Therefore the world's leading authorities in metrology developed a new concept of uncertainty in measurement. This concept is described in the Guide to the Expression of Uncertainty in Measurement (GUM) [1] and extended in the International Vocabulary of Metrology, third edition (VIM3) [2]. In accordance with the GUM and the VIM3, a result of measurement is generally expressed as a pair of values: a measured quantity value and its associated standard uncertainty. The measured value and the standard uncertainty together represent a range of values being attributed to the measurand [2, Sec. 2.9]. Suppose $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$ are n different results of measurement for a common measurand believed to be sufficiently stable, where x_1, …, x_n are the measured values and $u(x_1)$, …, $u(x_n)$ are the corresponding standard uncertainties. In the GUM concept of uncertainty, a measured value x_i and its associated standard uncertainty $u(x_i)$ are regarded, respectively, as the expected value and the standard deviation of an incompletely determined state-of-knowledge probability density function (pdf) attributed to the common measurand, for $i = 1, 2, \ldots, n$ [1].

Since the era of error analysis, metrologists have used the Birge chi-square test of statistical consistency to decide whether the differences between two or more measured values $x_1, \ldots, x_n$ are insignificant (Fig. 1). The Birge test is based on regarding the measured values $x_1, \ldots, x_n$ as realizations of random variables drawn from normal (Gaussian) sampling pdfs with unknown but equal expected values and known standard deviations [3]. When the measured values are correlated they are regarded as realizations of a random vector drawn from a joint n-variate normal distribution with a known variance-covariance matrix, referred to a normal consistency model. To assess statistical consistency of a set of measured values $x_1, \ldots, x_n$, a common practice is to pretend that the standard uncertainties $u(x_1), \ldots, u(x_n)$ are the known standard deviations of the presumed normal sampling pdfs of $x_1, \ldots, x_n$. It has previously been pointed out [4] that the Birge test and the concept of statistical consistency motivated by it do not apply to the results of measurement based on the GUM.

Recently, the VIM3 [2] introduced the idea of metrological compatibility, which can be used to assess the significance of the differences between two or more results of measurement for the same measurand (Fig. 2)**.** As noted in [4] the concept of metrological compatibility fits with the GUM and it can be used to assess the significance of differences between results based on the GUM for the same measurand. In Sec. 2, we discuss the VIM3 definition of metrological compatibility and its consequences in more detail than done in [4]. In this paper we propose an approach for determining a combined result which is metrologically compatible with the contributing results whether or not the results as available were compatible. When a set of results for the same measurand turn out to be incompatible, the seemingly anomalous results must be investigated. In Sec. 3, we discuss the importance of

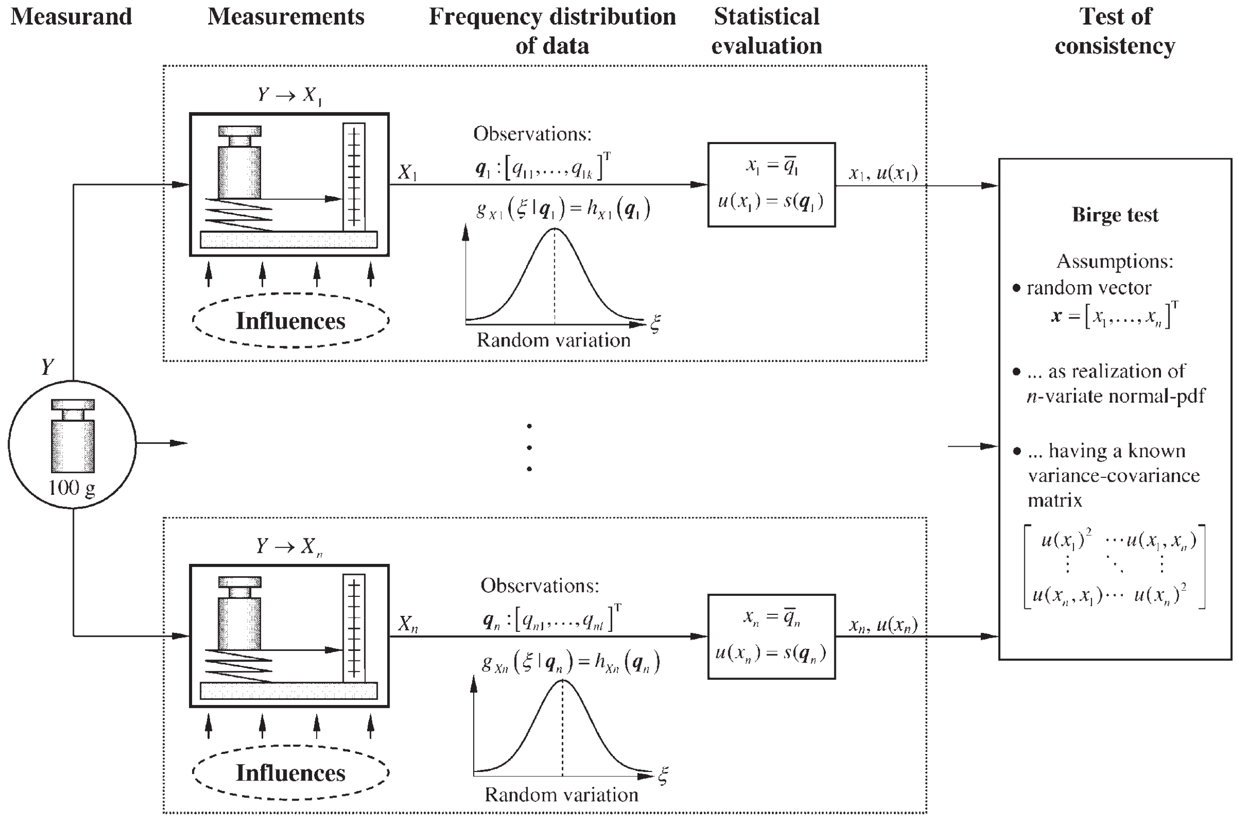
Measurand
Measurements
Frequency distribution of data
Statistical evaluation
Test of consistency
$Y \rightarrow X_1$
X_1
Observations:
$\boldsymbol{q}_1 : [q_{11}, \ldots, q_{1k}]^{\mathrm{T}}$
$g_{X1}(\xi \mid \boldsymbol{q}_1) = h_{X1}(\boldsymbol{q}_1)$
$x_1 = \bar{q}_1$
$u(x_1) = s(\boldsymbol{q}_1)$
$x_1, u(x_1)$
Influences
Random variation
ξ
Y
100 g
$Y \rightarrow X_n$
X_n
Observations:
$\boldsymbol{q}_n : [q_{n1}, \ldots, q_{nl}]^{\mathrm{T}}$
$g_{Xn}(\xi \mid \boldsymbol{q}_n) = h_{Xn}(\boldsymbol{q}_n)$
$x_n = \bar{q}_n$
$u(x_n) = s(\boldsymbol{q}_n)$
$x_n, u(x_n)$
Influences
Random variation
ξ
Birge test
Assumptions:
• random vector
$\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$
• ... as realization of n-variate normal-pdf
• ... having a known variance-covariance matrix
$\begin{bmatrix} u(x_1)^2 & \cdots u(x_1, x_n) \\ \vdots & \ddots & \vdots \\ u(x_n, x_1) \cdots & u(x_n)^2 \end{bmatrix}$

Fig. 1. Illustration of the classical approach to statistically evaluating and testing consistency of multiple measurements of the same measurand presuming a randomly disturbed measurement process. Symbols: Y – (joint) measurand, X_i – indicated quantities, ξ – possible values of the quantities X_i, $\boldsymbol{q}_1,\ldots,\boldsymbol{q}_n$ – vectors of the repeated observations q_{ij} where $\boldsymbol{q}_i = [q_{i1}, \ldots, q_{ik}]^{\mathrm{T}}$, $g_{Xi}(\xi \mid \boldsymbol{q}_i)$ – pdf for the quantity X_i given the data $\boldsymbol{q}_i$, $h_{Xi}(\boldsymbol{q}_i)$ – frequency distribution of the data $\boldsymbol{q}_i$.

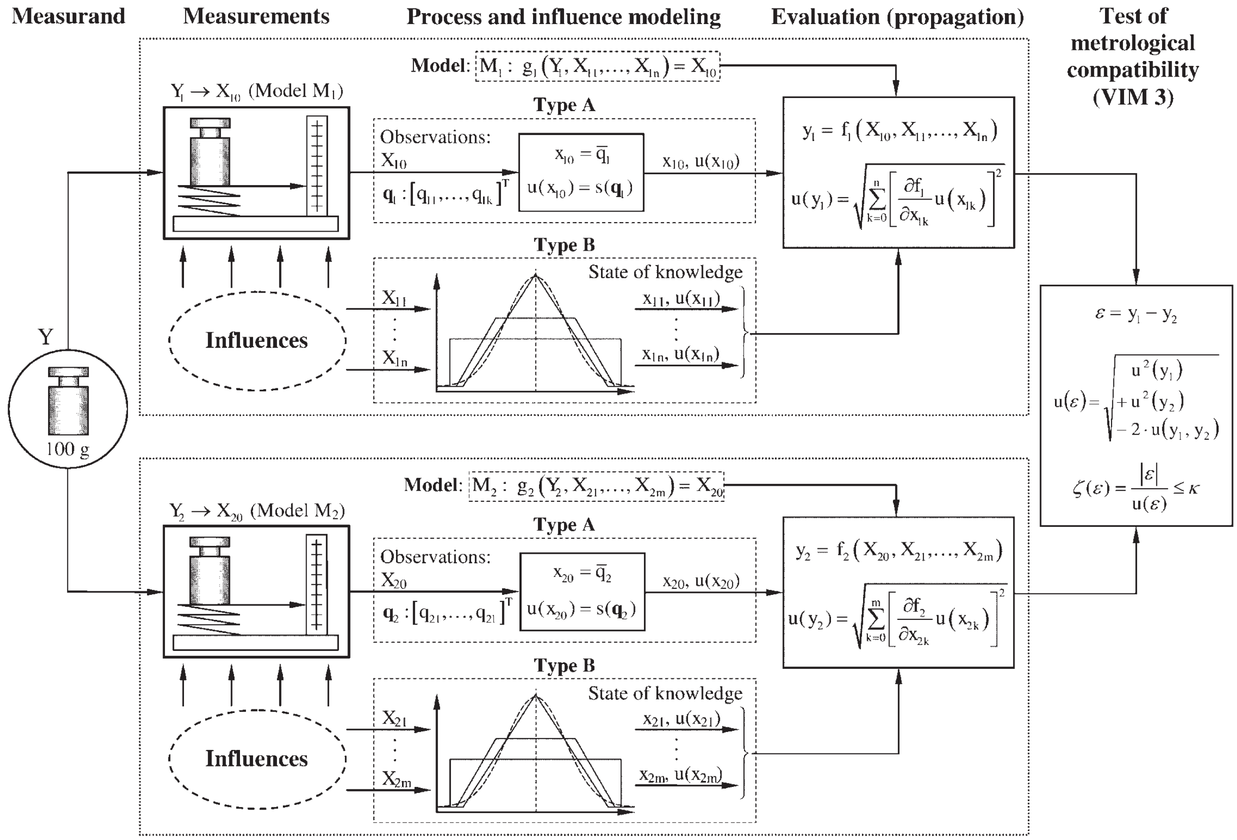
Measurand
Measurements
Process and influence modeling
Evaluation (propagation)
Test of metrological compatibility (VIM 3)
Model: $M_1: g_1(Y_1, X_{11}, \ldots, X_{1n}) = X_{10}$
$Y_1 \rightarrow X_{10}$ (Model M_1)
Type A
Observations: X_{10}
$\mathbf{q}_1 : [q_{11}, \ldots, q_{1k}]^T$
$x_{10} = \bar{q}_1$
$u(x_{10}) = s(\mathbf{q}_1)$
$x_{10}, u(x_{10})$
$y_1 = f_1(X_{10}, X_{11}, \ldots, X_{1n})$
$u(y_1) = \sqrt{\sum_{k=0}^{n} \left[\frac{\partial f_1}{\partial x_{1k}} u(x_{1k}) \right]^2}$
Type B
State of knowledge
X_{11}
X_{1n}
$x_{11}, u(x_{11})$
$x_{1n}, u(x_{1n})$
Influences
Y
100 g
$\varepsilon = y_1 - y_2$
$u(\varepsilon) = \sqrt{u^2(y_1) + u^2(y_2) - 2 \cdot u(y_1, y_2)}$
$\zeta(\varepsilon) = \frac{|\varepsilon|}{u(\varepsilon)} \leq \kappa$
Model: $M_2: g_2(Y_2, X_{21}, \ldots, X_{2m}) = X_{20}$
$Y_2 \rightarrow X_{20}$ (Model M_2)
Type A
Observations: X_{20}
$\mathbf{q}_2 : [q_{21}, \ldots, q_{2l}]^T$
$x_{20} = \bar{q}_2$
$u(x_{20}) = s(\mathbf{q}_2)$
$x_{20}, u(x_{20})$
$y_2 = f_2(X_{20}, X_{21}, \ldots, X_{2m})$
$u(y_2) = \sqrt{\sum_{k=0}^{m} \left[\frac{\partial f_2}{\partial x_{2k}} u(x_{2k}) \right]^2}$
Type B
State of knowledge
X_{21}
X_{2m}
$x_{21}, u(x_{21})$
$x_{2m}, u(x_{2m})$
Influences

Fig. 2. Illustration of an uncertainty approach to the metrological evaluation and test of compatibility of two measurements of the same measurand taking all known influences on the measurement processes into consideration. Symbols: Y – (joint) measurand, Y_1, Y_2 – measurand of the measurements, X_{10}, X_{20}, – indicated quantities, $\boldsymbol{q}_1$,$\boldsymbol{q}_2$ – vectors of the repeated observations q_{ij} where $\boldsymbol{q}_i = [q_{i1}, \ldots, q_{ik}]^T$, X_{1i}, X_{2i} – influence quantities with state of knowledge distributions. The elements in the grey blocks have been introduced by the GUM.

documenting information which may be needed in such investigations. Sometimes multiple evaluations of the same measurand need to be combined. A legitimate combined result must be metrologically compatible with the contributing results. In Sec. 4, we propose an approach for determining a combined result which is metrologically compatible with the contributing results. In Sec. 5, we illustrate the proposed approach using published data from an interlaboratory evaluation of the same measurand. A brief summary is given in Sec. 6.

2. The VIM3 Concept of Metrological Compatibility

Generally, the measurand (quantity intended to be measured) is a property of a material or of a phenomenon. In many scientific, industrial, and commercial measurements, the measurand is sufficiently stable between multiple evaluations. Our primary interest is in such applications. Suppose two or more measurement procedures are used to measure the same measurand. The measurement procedures may be (i) applications of the same method of measurement at different times or (ii) different implementations of a given method in different places or (iii) different methods.

A measured quantity value is a number together with a metrological reference (unit of measurement) expressing the magnitude of the quantity [2, Secs. 1.19 and 2.10] relative to the reference. A measured value must be traceable to a recognized metrological reference for it to be widely communicable. According to VIM3, two or more results of measurement for the same measurand are metrologically comparable if they are metrologically traceable to the same metrological reference [2, Sec. 2.46]. Metrological comparability does not imply that the measured values have similar magnitudes. The VIM3 concept of metrological compatibility applies only to those results of measurement which are metrologically comparable.

We assume that all results $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ for a common measurand are traceable to the same metrological reference and hence they are metrologically comparable. Following the GUM [1], we use the symbol X_i for a variable with a state-of-knowledge pdf represented by the result $[x_i, u(x_i)]$, for $i = 1, 2, \ldots, n$. The measured value x_i is regarded as the expected value $E(X_i)$ and the standard uncertainty $u(x_i)$ is regarded as the standard deviation $S(X_i)$ of the pdf of X_i for $i = 1, 2, \ldots, n$. In the mainstream GUM, the pdf of X_i is incompletely determined; the only thing reliably known about the pdf of X_i is the expected value $E(X_i) = x_i$ and the standard deviation $S(X_i) = u(x_i)$, for $i = 1, 2, \ldots, n$.

2.1 Metrological Compatibility of Two Particular Results

Metrological compatibility is defined for two results at a time. In the mainstream GUM, the difference $X_1 - X_2$ is a variable with an incompletely determined state-of-knowledge pdf for the difference between the values attributed by the two results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ to the common measurand. The expected value and the standard deviation of the pdf of $X_1 - X_2$ are, respectively, $E(X_1 - X_2) = x_1 - x_2$ and $S(X_1 - X_2) = \sqrt{[u^2(x_1) + u^2(x_2) - 2r(x_1, x_2)u(x_1)u(x_2)]}$, where $r(x_1, x_2)$ is the correlation coefficient between X_1 and X_2. Following the GUM, we use the symbol $u(x_1 - x_2)$ for the standard deviation $S(X_1 - X_2)$.

According to the VIM3 [2, Sec. 2.47], two metrologically comparable results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ for a measurand, supposed to be stable, are metrologically compatible if $|x_1 - x_2| \le \kappa \times u(x_1 - x_2)$ for a chosen threshold κ. According to the VIM3 [2, Sec. 2.47, Note 1], if two measurements for a common measurand, thought to be constant, are not metrologically compatible then there are two possibilities: (i) one or both of the measurements are incorrect (e.g., one or both of the measurement uncertainties are assessed as being too small) or (ii) the measurand changed between measurements.

We can use the VIM3 concept of metrological compatibility as a criterion to assess the significance of the differences between metrologically comparable results of measurement for the same measurand. In the mainstream GUM, the state-of-knowledge pdf represented by a result $[x_i, u(x_i)]$, for $i = 1, 2, \ldots, n$, is incompletely determined. Therefore, we need a quantitative measure for the difference between two fixed known results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$, each consisting of a measured value with standard uncertainty. Let us define a ζ-function, denoted by $\zeta(\Delta)$, as

$$\zeta(\Delta) = \frac{|\Delta|}{u(\Delta)} \,. \tag{1}$$

The value $\zeta(\Delta)$ is a measure for the significance of the difference Δ. Even when a complete state-of-knowledge pdf of Δ is assumed, the metric (1) can be used to judge on the significance of the difference. Based on this metric we can restate the VIM3 definition of metrological comparability as follows [4]:

Definition: Two metrologically comparable results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ for the same measurand are said to be metrologically compatible if

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{u(x_1 - x_2)} \le \kappa \,, \tag{2}$$

for a chosen value of some threshold κ, where

$$u(x_1 - x_2) = \sqrt{u^2(x_1) + u^2(x_2) - 2\,r(x_1, x_2)\,u(x_1)\,u(x_2)}, \tag{3}$$

and $r(x_1, x_2)$ is the correlation coefficient between the variables X_1 and X_2 with state-of-knowledge pdfs represented by the results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$.

In definition 1, the value of κ is a chosen threshold for declaring metrological compatibility (lack of significant difference) of two results. Values for $\zeta(x_1 – x_2)$ larger than κ are regarded as significant. The results are compatible, when the difference between the measured values x_1 and x_2 is insignificant in view of the standard uncertainties $u(x_1)$ and $u(x_2)$.

The VIM3 does not discuss how the threshold κ should be determined. A proper choice of the threshold κ is to a large extent a matter of agreement because it requires accepting the economic consequences of that choice. A conventional value of the threshold κ in metrology is two.

If one would agree on a larger value for κ then small differences are not detectable any more. This would be a disadvantage for applications when detecting small differences is important. But if we would agree on a smaller value for κ then a lot of small differences become significant even though they might be only a consequence of noisy measurements and the economic consequences are suffered by the metrological community trying to provide compatible measurement systems.

2.2 Metrological Compatibility of a Set of Results

According to the VIM3 [2, Sec. 2.47], a set of comparable results $[x_1, u(x_1)]$, …, $[x_n, u(x_n)]$, where $n \geq 2$, is metrologically compatible if every one of the $n(n-1)/2$ pairs of results $[x_i, u(x_i)]$ and $[x_j, u(x_j)]$, for $i, j = 1, 2, …, n$ and $i < j$, is metrologically compatible. We can use expression (2) in this case by replacing x_1 with x_i and x_2 with x_j.

If for all pairs of results the values of $\zeta\,(x_i - x_j)$ are smaller than or equal to a chosen threshold κ then the set of results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, …, $[x_n, u(x_n)]$ is metrologically compatible.

We can say that the differences between the measured values x_1, …, x_n are insignificant in view of the uncertainties $u(x_1)$, …, $u(x_n)$.

Note 1: A conventional idea that if the number n of the measured values x_1, …, x_n is large, it is natural to expect one or more of them to be significantly different from the rest comes from the theory of sampling from probability distributions having long tails which extend, for example, beyond two standard deviations. If the measurement procedures are properly carried out and the results of measurement are properly evaluated according to the GUM taking into account all important influence quantities, then a set of results for the same measurand should be metrologically compatible. When some results of measurement seem anomalous, they require explanation rather than acceptance. Often, anomalous results are consequence of missing important influence quantities.

2.3 Metrological Compatibility With a Reference Result

Suppose that in addition to the n measurement procedures, which yield the comparable results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, …, $[x_n, u(x_n)]$, where $n \geq 2$, the same measurand is measured by a higher echelon measurement procedure (or laboratory) yielding the reference result $[x_R, u(x_R)]$, where x_R is the reference value with standard uncertainty $u(x_R)$. Alternatively, the common measurand may be a certified reference material of reference value x_R with standard uncertainty $u(x_R)$, which are not revealed before all n results of measurement are reported. We will use the symbol X_R for a variable with a state-of-knowledge pdf represented by the result $[x_R, u(x_R)]$. In general, the uncertainty $u(x_R)$ associated with the reference value x_R is smaller than the uncertainties $u(x_1)$, …, $u(x_n)$ associated with the measured values x_1, …, x_n.

If for all differences between the results x_i and value x_R, the values $\zeta\,(x_i - x_R)$ are smaller than or equal to a chosen threshold κ then the set of results $[x_1, u(x_1)]$, $[x_2, u(x_{2)}]$, …, $[x_n, u(x_n)]$ is metrologically compatible with the reference value x_R. We can say that the differences between the measured values x_1, …, x_n and the reference value x_R are insignificant in view of the uncertainties $u(x_1)$, …, $u(x_n)$ and $u(x_R)$.

One should not confuse the difference $\zeta\,(x_i - x_R)$ between the results $[x_i, u(x_i)]$ and $[x_R, u(x_R)]$ with E_n-values which do not seem to be uniquely defined.[1]

2.4 Metrological Compatibility With a Combined Result

Sometimes the results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, …, $[x_n, u(\mathrm{x}_n)]$, where $n \geq 2$, need to be combined to determine a combined result $[x_C, u(x_C)]$, where x_C is the combined value and $u(x_C)$ is the standard uncertainty associated with x_C. We will use the symbol X_C for a variable with a state-of-knowledge pdf represented by $[x_C, u(x_C)]$. In accordance with the GUM, the combined variable X_C for a value of the measurand should be defined as a measurement function of the input variables X_1, …, X_n. Often, X_C is set as a convex linear combination of X_1, …, X_n with non-negative weights a_1, …, a_n which sum up to one. Thus often a measurement function for X_C is of the form

$$X_C = \sum_i a_i X_i \,, \tag{4}$$

where $a_i \geq 0$ and $\Sigma_i a_i = 1$, for $i = 1, 2, …, n$. Since (4) is a linear function in X_i the expected value $E(X_C)$ of X_C is the combined value x_C, where

$$X_C = \sum_i a_i x_i \,, \tag{5}$$

and the standard deviation $S(X_C)$ of X_C is the standard uncertainty $u(x_C)$ where

$$\begin{aligned} u^2(x_C) &= \sum_i a_i^2 u^2(x_i) \\ &\quad + 2\sum_{i<j} a_i a_j u(x_i) u(x_j) r(x_i, x_j) \,. \end{aligned} \tag{6}$$

[1] One version defines E_n-value as $E_n = (x_i - x_R)/\sqrt{[(2s_i)^2 + (2s_R)^2]}$, where x_i and x_R are regarded as realizations of random variables with sampling pdfs and s_i and s_R are the estimated standard deviations of those sampling pdfs. Thus E_n-values are realizations of random variables with sampling pdfs. Some metrologists substitute in the denominator of the E_n-value, the expanded standard uncertainties for $2s_i$ and $2s_R$. This is inappropriate uses of the expanded standard uncertainties.

If the individual measurement procedures are all uncorrelated then the cross-product term in (6) is zero.

If $a_i = 1/n$ for $i = 1, 2, \ldots, n$, then X_{C} reduces to the arithmetic average $X_{\mathrm{A}} = (1/n)\ \Sigma_i X_i$. The expected value $E(X_{\mathrm{A}})$ is $x_{\mathrm{A}} = (1/n)\ \Sigma_i x_i$ and the standard deviation $S(X_{\mathrm{A}})$ denoted by $u(x_{\mathrm{A}})$ can be determined from (6). If the pdfs for $X_1, \ldots, X_n$ are uncorrelated, then

$$u^2(x_{\mathrm{A}}) = \frac{1}{n^2} \sum\nolimits_i u^2(x_i)\,. \qquad (7)$$

If $a_i = w_i / \Sigma_i w_i$, where $w_i = 1/u^2(x_i)$ then X_{C} reduces to the weighted mean $X_{\mathrm{W}} = \Sigma_i w_i X_i / \Sigma_i w_i$ with weights inversely proportional to the variances $u^2(x_1), \ldots, u^2(x_n)$. The expected value $E(X_{\mathrm{W}})$ is $x_{\mathrm{W}} = \Sigma_i w_i x_i / \Sigma_i w_i$ and the standard deviation $S(X_{\mathrm{W}})$ denoted by $u(x_{\mathrm{W}})$ can be determined from (6). If the pdfs for $X_1, \ldots, X_n$ are uncorrelated, then

$$u^2(x_{\mathrm{W}}) = \frac{1}{\Sigma_i w_i} = \frac{1}{\Sigma_i (1/u^2(x_i))}\,. \qquad (8)$$

If for all differences between the results x_i and the combined value x_{C}, the values $\zeta(x_i - x_{\mathrm{C}})$ are smaller than or equal to a chosen threshold κ then the set of results $[x_1, u(x_1)]$, $[x_2, u(x_2)]$, $\ldots$, $[x_n, u(x_n)]$ is metrologically compatible with the combined value x_{C}. Then we can say that the differences between the measured values $x_1, \ldots, x_n$ and the combined value x_{C} are insignificant in view of the uncertainties $u(x_1), \ldots, u(x_n)$.

In evaluating $u(x_i - x_{\mathrm{C}})$ the correlation coefficient between X_i and X_{C} must be included because the pdfs of X_i and X_{C} are always correlated, for $i = 1, 2, \ldots, n$. For example, if the pdfs for $X_1, \ldots, X_n$ are uncorrelated, then the variance, $V(X_i - X_{\mathrm{C}})$, denoted by $u^2(x_i - x_{\mathrm{C}})$ is

$$u^2(x_i - x_{\mathrm{C}}) = u^2(x_i) + \sum\nolimits_i a_i^2\, u^2(x_i) - 2\, a_i\, u^2(x_i)\,. \qquad (9)$$

If $a_i = 1/n$, for $i = 1, 2, \ldots, n$, then x_{C} reduces to the arithmetic average $x_{\mathrm{A}} = (1/n)\ \Sigma_i x_i$ and the uncertainty $u(x_i - x_{\mathrm{C}})$ given in (9) reduces to $u(x_i - x_{\mathrm{A}})$, where

$$u^2(x_i - x_{\mathrm{A}}) = \left(\frac{n-2}{n}\right) u^2(x_i) + u^2(x_{\mathrm{A}})\,. \qquad (10)$$

If $a_i = w_i / \Sigma_i w_i$, where $w_i = 1/u^2(x_i)$, for $i = 1, 2, \ldots, n$, then x_{C} reduces to the weighted mean $x_{\mathrm{W}} = \Sigma_i w_i x_i / \Sigma_i w_i$ and the uncertainty $u(x_i - x_{\mathrm{C}})$ given in (9) reduces to $u(x_i - x_{\mathrm{W}})$, where

$$u^2(x_i - x_{\mathrm{W}}) = u^2(x_i) - u^2(x_{\mathrm{W}})\,. \qquad (11)$$

If the uncertainties $u(x_1), u(x_2), \ldots, u(x_n)$ were all equal to $u(x)$, say, then x_{W} reduces to x_{A} and $u^2(x_{\mathrm{W}})$ reduces to $u^2(x_{\mathrm{A}}) = u^2(x)/n$. Then both (10) and (11) reduce to

$$u^2(x) - \frac{u^2(x)}{n} = \frac{n-1}{n}\, u^2(x)\,. \qquad (12)$$

Note 2: Sometimes, the standard uncertainties $u(x_1)$, $u(x_2), \ldots, u(x_n)$ are not all reliably determined. Also, the standard uncertainties are frequently inappropriate bases for assigning the weights $a_1, a_2, \ldots, a_n$ to the measured values $x_1, x_2, \ldots, x_n$ to determine a combined result. Therefore the weighted mean x_{W} may be inappropriate for combining the values. Thus, in our view, the arithmetic mean x_{A} should be regarded as a default combined value.

3. Information Needed to Determine Sources of Incompatibility

A purpose of assessing metrological compatibility is to demonstrate lack of significant difference between the results of measurement for a common measurand. If a set of results turns out to be metrologically incompatible then the measurement procedures and calculations underlying the seemingly anomalous results should be investigated. Every result of measurement should have supporting documents which include the measurement function (measurement equation) and complete uncertainty budget. If the influence quantities, uncertainty components, and correlation coefficients identified in the uncertainty budget are reasonable then in search of the possible sources of incompatibility one must look into potential influence quantities not included in the uncertainty budget.

Investigations to determine the sources of incompatibility are generally done in retrospect long after completing the measurements. Therefore investigators need detailed descriptions of what was actually done during measurement. Often, metrologists do not have enough time and resources to document in sufficient detail for retrospective investigation what was actually done in a particular application of the measurement procedure. In the absence of such documentation it may be difficult to determine possible sources of incompatibility.

Note 3: We hope that in the not too distant future, metrologists and information technology experts would collaborate to develop tools which make it easier for metrologists to document in real time the actual measurement procedure while the measurements are being done. Such documentation should be helpful in identifying all potentially important influence quantities.

4. Determination of a Combined Value and Its Associated Uncertainty

Even when the common measurand is sufficiently stable, the results $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ can exhibit large variation. Metrological incompatibility occurs when some or all results (measured values or standard uncertainties) are improperly determined. Frequently, improper results are consequence of missing important influence quantities. For example, in many chemical measurements, the measurand is the amount of one component in a sample of multi-component material. The other components can interfere with the measurements. Frequently, it is impossible to know all potential interferences. Therefore, it is difficult to be sure that all significant influence quantities have been accounted for in determining the measured values and uncertainties.

For a combined result $[x_\mathrm{C}, u(x_\mathrm{C})]$ to be legitimate it should be metrologically compatible with the contributing results of measurement $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. Therefore we propose the following principle.

Principle for combining multiple results for the same measurand: Determine the combined result $[x_\mathrm{C}, u(x_\mathrm{C})]$ from the expressions (5) and (6) as recommended in the GUM. If the results $[x_1, u(x_1)], [x_2, u(x_2)], \ldots, [x_n, u(x_n)]$ are metrologically compatible with the combined result $[x_\mathrm{C}, u(x_\mathrm{C})]$, then $u(x_\mathrm{C})$ is a valid expression for the standard uncertainty associated with x_C. If the results $[x_1, u(x_1)], [x_2, u(x_2)], \ldots, [x_n, u(x_n)]$ are metrologically incompatible with the combined result $[x_\mathrm{C}, u(x_\mathrm{C})]$, then the seemingly anomalous results should be investigated. Until the investigation resolves the anomalous results, in the absence of additional knowledge, all results in a metrologically incompatible set should be regarded with suspicion. To determine a legitimate combined result, we propose that the measured values $x_1, \ldots, x_n$ should be sustained and each of the uncertainties $u(x_1), u(x_2), \ldots, u(x_n)$ should be enlarged just enough to make the results $[x_1, u(x_1)], [x_2, u(x_2)], \ldots, [x_n, u(x_n)]$ metrologically compatible with the combined result $[x_\mathrm{C}, u(x_\mathrm{C})]$.

This approach was first proposed in [5] and has recently been used in [6]. Thus we define variables $Y_1, \ldots, Y_n$ with corrected state-of-knowledge pdfs for the common measurand as follows

$$Y_i = X_i + \delta X_i \ , \tag{13}$$

where $\delta X_1, \ldots, \delta X_n$ are correction variables. Then a measurement function for the combined variable Y_C is

$$\begin{aligned} Y_\mathrm{C} &= \sum\nolimits_i a_i Y_i = \sum\nolimits_i a_i X_i + \sum\nolimits_i a_i \delta X_i \\ &= X_C + \sum\nolimits_i a_i \delta X_i \ , \end{aligned} \tag{14}$$

where $a_i \geq 0$ and $\Sigma_i\, a_i = 1$, for $i = 1, 2, \ldots, n$, and the pdfs for the correction variables $\delta X_1, \ldots, \delta X_n$ are mutually independent and independent of the pdfs for $X_1, \ldots, X_n$. The pdfs assigned to the correction variables $\delta X_1, \ldots, \delta X_n$ express the limits of knowledge. Thus, we assign zero expected values and the same variance $u^2(\delta)$ to each of the correction variables $\delta X_1, \ldots, \delta X_n$. Thus the expected value $E(\delta X_i)$ is zero and the variance $V(\delta X_i)$ is $u^2(\delta)$, for $i = 1, 2, \ldots, n$. It follows from (13) that the expected value y_i and the variance $u^2(y_i)$ of the pdf for Y_i are

$$y_i = x_i + 0 = x_i \ , \tag{15}$$

and

$$u^2(y_i) = u^2(x_i) + u^2(\delta) \ , \tag{16}$$

for $i = 1, 2, \ldots, n$.

We propose that the variance $u^2(\delta)$ should be set just large enough to make the results $[y_1, u(y_1)], [y_2, u(y_2)], \ldots, [y_n, u(y_n)]$ compatible with the result $[y_\mathrm{C}, u(y_\mathrm{C})]$. As discussed in Sec. 2.4, the results $[y_1, u(y_1)], [y_2, u(y_2)], \ldots, [y_n, u(y_n)]$ are compatible with $[y_\mathrm{C}, u(y_\mathrm{C})]$ when

$$\zeta(y_i - y_\mathrm{C}) = \frac{|y_i - y_\mathrm{C}|}{u(y_i - y_\mathrm{C}} \leq \kappa \ , \tag{17}$$

or equivalently

$$(y_i - y_\mathrm{C})^2 \leq \kappa^2 \times u^2(y_i - y_\mathrm{C}) \ , \tag{18}$$

for all $i = 1, 2, \ldots, n$. From (15), we have

$$y_\mathrm{C} = \sum\nolimits_i a_i y_i = \sum\nolimits_i a_i x_i = x_\mathrm{C} \, , \tag{19}$$

and

$$y_i - y_\mathrm{C} = x_i - x_\mathrm{C} \ . \tag{20}$$

From the appendix, we have

$$u^2(y_i - y_\mathrm{C}) = u^2(x_i - x_\mathrm{C}) + u^2(\delta)[1 + \sum\nolimits_i a_i^2 - 2a_i] . \tag{21}$$

Therefore, the criterion of compatibility (18) is equivalent to

$$(x_i - x_\mathrm{C})^2 \le \kappa^2 \times \left(u^2(x_i - x_\mathrm{C}) + u^2(\delta)[1 + \sum\nolimits_i a_i^2 - 2a_i]\right), \tag{22}$$

for all $i = 1, 2, \ldots, n$. It follows that

$$u^2(\delta) \ge \frac{1}{[1 + \sum_i a_i^2 - 2a_i]} \left(\frac{(x_i - x_\mathrm{C})^2}{\kappa^2} - u^2(x_i - x_\mathrm{C}) \right), \tag{23}$$

for all $i = 1, 2, \ldots, n$. Thus, if $u^2(\delta)$ is chosen as

$$u^2(\delta) = \max\left(0, \left\{ \frac{1}{[1 + \sum_i a_i^2 - 2a_i]} \left(\frac{(x_i - x_\mathrm{C})^2}{\kappa^2} - u^2(x_i - x_\mathrm{C}) \right) \middle| \; i = 1, \ldots, n \right\}\right), \tag{24}$$

then each of the corrected measured values $y_1, \ldots, y_n$ would be metrologically compatible with the combined measured value y_C. If the measured values $x_1, \ldots, x_n$ are compatible with the combined measured value x_C then each of the n quantities in the curly parenthesis of (24) are negative and $u^2(\delta) = 0$. In that case the measurement function (14) reduces to (5) and the uncertainty associated with the combined measured value x_C is given by (6).

4.1 Arithmetic Average

If $a_i = 1/n$, for $i = 1, 2, \ldots, n$, then x_C reduces to the arithmetic average x_A and from (24),

$$u^2(\delta) = \max\left(0, \left\{ \frac{n}{n-1} \left(\frac{(x_i - x_\mathrm{A})^2}{\kappa^2} - u^2(x_i - x_\mathrm{A}) \right) \middle| \; i = 1, \ldots, n \right\}\right). \tag{25}$$

The combined value y_C reduces to $y_\mathrm{A} = (1/n)\, \Sigma_i\, y_i = (1/n)\, \Sigma_i\, x_i = x_\mathrm{A}$. To assure that the measured values $y_1, \ldots, y_n$ are compatible with y_A one can check that

$$\zeta(y_i - y_\mathrm{A}) = \frac{|y_i - y_\mathrm{A}|}{u(y_i - y_\mathrm{A})} \le \kappa , \tag{26}$$

where as shown in the appendix

$$u^2(y_i - y_\mathrm{A}) = u^2(x_i - x_\mathrm{A}) + \frac{n-1}{n} u^2(\delta) . \tag{27}$$

Expressions for $u^2(x_i - x_\mathrm{A})$ and $u^2(\delta)$ are given in (10) and (25), respectively. The uncertainty associated with y_A is from (7)

$$u^2(y_\mathrm{A}) = \frac{1}{n^2} \sum\nolimits_i u^2(y_i) = \frac{1}{n^2} \sum\nolimits_i (u^2(x_i) + u^2(\delta)) . \tag{28}$$

If $u^2(\delta) = 0$, then (28) reduces to (7).

4.2 Weighted Mean

Since the variance associated with y_i is $u^2(y_i) = u^2(x_i) + u^2(\delta)$, a weighted mean with weights inversely proportional to the variances of the results $y_1, \ldots, y_n$ is $y_\mathrm{W} = \Sigma_i\, w_i\, y_i / \Sigma_i\, w_i$, where $y_i = x_i$, and $w_i = 1/u^2(y_i) = 1/[u^2(x_i) + u^2(\delta)]$ for $i = 1, 2, \ldots, n$. The measured values $y_1, \ldots, y_n$ are compatible with y_W if

$$\zeta(y_i - y_\mathrm{W}) = \frac{|y_i - y_\mathrm{W}|}{u(y_i - y_\mathrm{W})} \le \kappa , \tag{29}$$

for all $i = 1, 2, \ldots, n$. Analogous to (11)

$$u^2(y_i - y_\mathrm{W}) = u^2(y_i) - u^2(y_\mathrm{W}) , \tag{30}$$

where

$$u^2(y_\mathrm{W}) = \frac{1}{\sum_i (1/[u^2(x_i) + u^2(\delta)])} . \tag{31}$$

The variance $u^2(\delta)$ is the smallest value which would make the measured values $y_1, \ldots, y_n$ compatible with y_W. Such a value for $u^2(\delta)$ can be iteratively determined using the value of $u^2(\delta)$ from (25) as a starting value.

Note 4: Let us use the symbol Y_{true} for a true quantity value [2, Sec. 2.11] of the common measurand commensurate with its description. (In the GUM, the same symbol Y is also used for a quantity with a state-of-knowledge pdf for the common measurand.) If the measurand is defined in extensive detail, a true value Y_{true} may be essentially unique. If the measurand is defined in less detail, then a range of values may be commensurate with its definition and any one of them qualifies as a true value Y_{true} of the measurand. The concept of metrological compatibility relates to the observed differences between the measured values $x_1, \ldots, x_n$ rather than to the unobservable differences between the measured values and a true value Y_{true} of the measurand. Therefore, regardless of whether the measured values $x_1, \ldots, x_n$ are compatible or incompatible with the combined value x_C, the measured values alone provide no information about the difference between x_C and Y_{true}. In particular, metrological compatibility does not imply that the difference between x_C and Y_{true} is not significant. However, there is no factual knowledge about potential significant difference between x_C and Y_{true}. Therefore, a correction applied to x_C for its potential significant difference between x_C and Y_{true} and enlargement of the uncertainty $u(x_C)$ determined from (6) as discussed in [7] would be arbitrary.

5. Combined Result From an Interlaboratory Evaluation

The Columns 2 and 3 of table 1 reproduce from [8, Table 3] the measured values, c_{Lab}, and the corresponding standard uncertainties, $u(c_{\text{Lab}})$, for the amount content of lead (Pb) in natural river water as determined by the eight laboratories[2] identified in column 1 of table 1. We will use these data to illustrate calculation of a combined result. Suppose the arithmetic average $c_{\text{Avg}} = 62.79$ nmol/kg is used as the combined measured value. The associated standard uncertainty based on the expression (7) is $u(c_{\text{Avg}}) = 0.26$ nmol/kg. The values of $\zeta(c_{\text{Lab}} - c_{\text{Avg}})$ between the reported results $[c_{\text{Lab}}, u(c_{\text{Lab}})]$ and the combined result $[c_{\text{Avg}}, u(c_{\text{Avg}})]$ determined by using the expression (10) for the standard uncertainty $u(c_{\text{Lab}} - c_{\text{Avg}})$ are shown in column 4 of table 1. Suppose the threshold for metrological compatibility is set as $\kappa = 2$. One of the values of $\zeta(c_{\text{Lab}} - c_{\text{Avg}})$ (from LNE) is larger than 2.00. Therefore not all of the eight reported results $[c_{\text{Lab}}, u(c_{\text{Lab}})]$ are metrologically compatible with the combined result $[c_{\text{Avg}}, u(c_{\text{Avg}})]$. Until potential flaws in the deviant result (from LNE or the others) are determined, all results must be regarded with suspicion. Therefore, as discussed in Sec. 4, we propose that all reported measured values should be sustained and each of the uncertainties should be enlarged by the amount $u^2(\delta) = 1.130$ determined from the expression (25). The adjusted (enlarged) standard uncertainties $u(c_{\text{Lab}})$ based on the expression (16) are shown in column 5 of table 1. Based on the adjusted uncertainties $u(c_{\text{Lab}})$, the standard uncertainty associated with the arithmetic mean c_{Avg} determined from the expression (28) is $u(c_{\text{Avg}}) = 0.46$ nmol/kg. The differences $\zeta(c_{\text{Lab}}, c_{\text{Avg}})$ based on the adjusted uncertainties are shown in column 6 of table 1. Since none of the values of $\zeta(c_{\text{Lab}} - c_{\text{Avg}})$ is larger than 2.00, the adjusted results $[c_{\text{Lab}}, u(c_{\text{Lab}})]$ given in columns 2 and 5 of table 1 are metrologically compatible with the combined result $[c_{\text{Avg}}, u(c_{\text{Avg}})]$.

Table 1. The measured values c_{Lab} for the amount content of Pb in natural river water and their associated standard uncertainties $u(c_{\text{Lab}})$ in nmol/kg units as reported in [8]. Also shown are the differences $\zeta(c_{\text{Lab}} - c_{\text{Avg}})$ based on the reported uncertainties and the adjusted (enlarged) uncertainties

Laboratory Identifier	Amount Content c_{Lab}/ nmol/kg	Reported Uncertainty $u(c_{\text{Lab}})$/ nmol/kg	Reported ζ-value	Adjusted Uncertainty $u(c_{\text{Lab}})$/ nmol/kg	Adjusted ζ-value
NMi	61.40	1.10	1.40	1.53	0.99
NIMC	62.21	0.30	1.56	1.10	0.54
KRISS	62.30	0.45	1.04	1.15	0.44
LGC	62.34	0.62	0.75	1.23	0.38
NRC	62.60	0.75	0.27	1.30	0.15
IRMM	62.70	0.26	0.25	1.09	0.08
NIST	62.84	0.15	0.19	1.07	0.05
LNE	65.90	1.35	2.60	1.72	2.00

Figures 3 and 4 display the measured values c_{Lab} (given in column 2 of table 1) and the arithmetic average $c_{\text{Avg}} = 62.79$ nmol/kg along with the corresponding expanded uncertainty intervals (for coverage factor $k = 2$). In Fig. 3, the expanded uncertainty intervals are based on the standard uncertainties as reported in [8] and reproduced in column 3 of table 1; in particular, the standard uncertainty $u(c_{\text{Avg}})$ associated with c_{Avg} is $u(c_{\text{Avg}}) = 0.26$ nmol/kg. In Fig. 4, the expanded uncertainty intervals are based on the adjusted (enlarged) standard uncertainties displayed in column 5 of table 1; in particular, the standard uncertainty $u(c_{\text{Avg}})$ associated with c_{Avg} is $u(c_{\text{Avg}}) = 0.46$ nmol/kg.

[2] Reference [8] is the final report of the CIPM international key comparison CCQM-K2. In this paper, we have used data from [8] to illustrate calculation of a combined result. We do not address data analysis of a key comparison to determine the key comparison reference value (KCRV) and the degrees of equivalence (DOE).

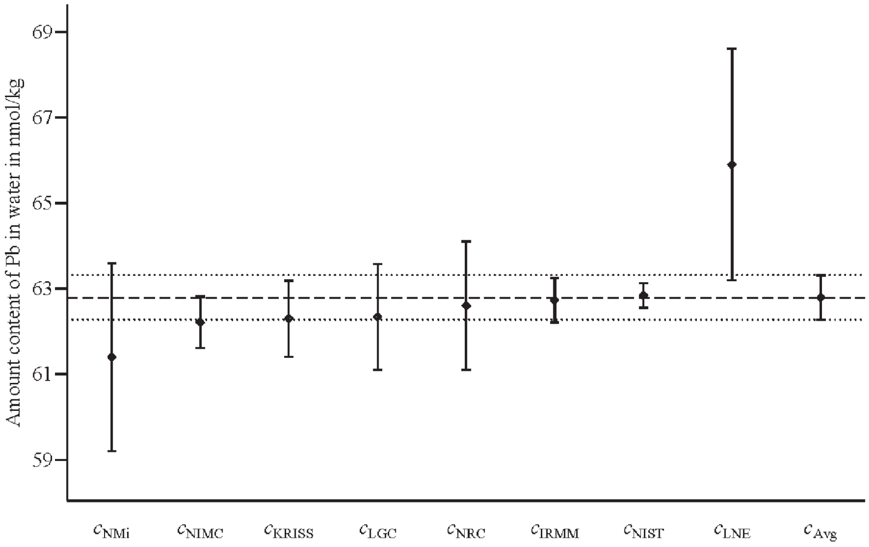

Fig. 3. The measured values c_{Lab} and their arithmetic average c_{Avg} for the amount content of lead (Pb) with the expanded uncertainty intervals (for coverage factor $k = 2$) determined from the uncertainties stated in the report [8] and reproduced in column 3 of table 1. The arithmetic average is $c_{Avg} = 62.79$ nmol / kg with standard uncertainty $u(c_{Avg}) = 0.26$ nmol / kg.

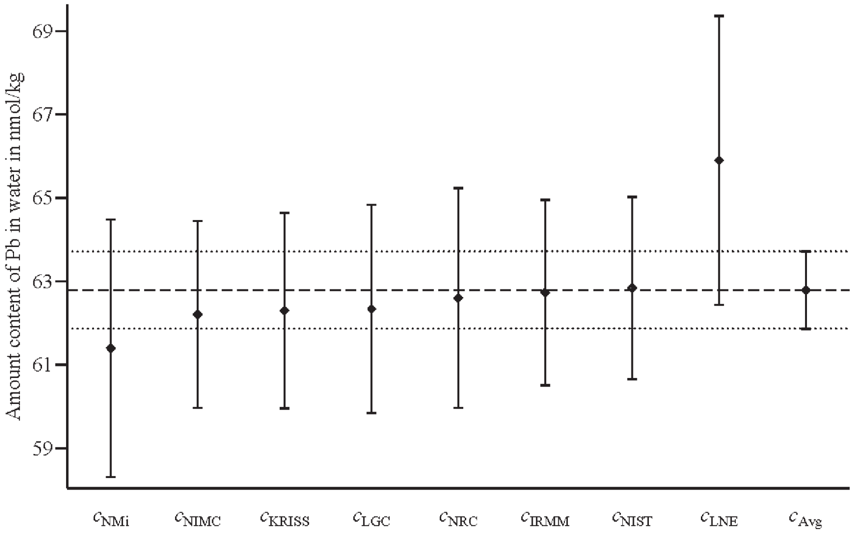

Fig. 4. The measured values c_{Lab} and their arithmetic average c_{Avg} for the amount content of lead (Pb) with the expanded uncertainty intervals (for coverage factor $k = 2$) determined from the adjusted (enlarged) uncertainties given in column 5 of table 1. The arithmetic average is $c_{Avg} = 62.79$ nmol / kg with standard uncertainty $u(c_{Avg}) = 0.46$ nmol / kg.

In both Figs. 3 and 4, the expanded uncertainty intervals (for coverage factor $k = 2$) for the measured values overlap with the expanded uncertainty interval for the arithmetic average c_{Avg}. However, not all of the eight results in Fig. 3 are metrologically compatible with the combined result [c_{Avg}, $u(c_{Avg})$]. This shows that there is no direct correspondence between the overlap of the expanded uncertainty intervals (for coverage factor $k = 2$) and the VIM3 concept of metrological compatibility.

6. Summary

The VIM3 [2] concept of metrological compatibility applies to only those results which are metrologically comparable; that is, the results must be traceable to the same reference. Metrological compatibility is a pairwise concept. Two metrologically comparable results for the same measurand are said to be metrologically compatible if the ζ-value of the difference between the results is less than or equal to a chosen threshold (usually 2.0). A set of metrologically comparable results is metrologically compatible if all of the distinct pairs of results are metrologically compatible. The concept of metrological compatibility easily extends to compatibility of a set of results with a reference result or a combined result. Metrological compatibility does not require complete knowledge of the pdfs represented by the results of measurement.

Often multiple evaluations for the same measurand must be combined to determine a combined result. For a combined result to be legitimate it should be metrologically compatible with the contributing results of measurement. When the results are metrologically incompatible with the combined result, we propose that the measured values should be sustained and each of the standard uncertainties should be enlarged just enough to make the results compatible with the combined result. Then the results can be combined using the GUM. This approach has been found to be useful in many practical applications.

7. Appendix

Since $\delta X_{\mathrm{C}} = \Sigma_i\, a_i \delta X_i$, we have expected value $E(\delta X_{\mathrm{C}}) = 0$ and variance $V(\delta X_{\mathrm{C}}) = u^2(\delta)\Sigma_i\, a_i^2$. Thus $E(\delta X_i - \delta X_{\mathrm{C}}) = 0$ and $V(\delta X_i - \delta X_{\mathrm{C}}) = V(\delta X_i) + V(\delta X_{\mathrm{C}}) - 2C(\delta X_i, \delta X_{\mathrm{C}}) = u^2(\delta)[1+\Sigma_i\, a_i^2 - 2a_i]$, where the third term is covariance. Therefore
$u^2(y_i - y_{\mathrm{C}}) = V(Y_i - Y_{\mathrm{C}}) = V(X_i - X_{\mathrm{C}}) + V(\delta X_i - \delta X_{\mathrm{C}}) = u^2(x_i - x_{\mathrm{C}}) + u^2(\delta)[1+\Sigma_i\, a_i^2 - 2a_i]$.

If $a_i = 1/n$, for $i = 1, 2, \ldots, n$, then $u^2(y_i - y_{\mathrm{A}}) = u^2(x_i - x_{\mathrm{A}}) + \dfrac{n-1}{n} u^2(\delta)$.

Acknowledgment

We thank Javier Bernal, Tyler Estler and two anonymous referees for their comments on earlier drafts of this paper.

8. References

[1] ISO 1995 Guide to the Expression of Uncertainty in Measurement (GUM), 2nd ed., (Geneva: International Organization for Standardization).

[2] BIPM/JCGM International Vocabulary of Metrology— Basic and general concepts and associated terms. 3rd ed., (Sèvres: Bureau International des Poids et Mesures, Joint Committee for Guides in Metrology) (2008). (http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf)

[3] R. N. Kacker, A. B. Forbes, R. Kessel, and K. Sommer, Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations, Metrologia **45**, 257-264 (2008).

[4] R. N. Kacker, R. Kessel, and K. Sommer, Assessing differences between results determined according to the Guide to the Expression of Uncertainty in Measurement, J. Res. Natl. Stand. Technol. **115**, 453-459 (2010).

[5] R. Kessel, M. Bergland, and R.Wellum, Application of consistency checking to evaluation of uncertainty in multiple replicate measurements Accreditation and Quality Assurance: Journal of Quality, Comparability and Reliability in Chemical Measurement **13**, 293-298 (2008).

[6] R. Wellum, A. Verbruggen, and R. Kessel, A new evaluation of the half-life of ^{241}Pu, J. Analytical Atomic Spectrometry **24**, 801-807 (2009).

[7] R. N. Kacker, R. U. Datla, and A. C. Parr, Statistical analysis of CIPM key comparisons based on the ISO Guide, Metrologia **41**, 340-352 (2004).

[8] I. Papadakis, P. D. P. Taylor, and P. De Bièvre, CCQM-K2 key comparison: cadmium and lead content in natural water, Metrologia **38**, 543-547 (2001).

About the authors**:** *Raghu Kacker is a researcher in the Applied and Computational Mathematics Division (ACMD) of the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST). His current interests include software testing and evaluation of the uncertainty in outputs of computational models and physical measurements. He has co-authored over 100 refereed papers. He has a Ph.D. in statistics. He is a Fellow of the American Statistical Association and a Fellow of the American Society for Quality. Rüdiger Kessel was a guest researcher in the Applied and Computational Mathematics Division (ACMD) of the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST). He is an electronic and data systems engineer and currently a researcher at Physikalisch Technische Bundesanstalt in*

Germany. He has a Ph.D. in sciences from the Analytical Chemistry Department of the University of Antwerp, Belgium and he is the developer of a standard software tool to evaluate uncertainty of measurement. His current interests include evaluation of uncertainty in physical and chemical measurements, modelling of measurements and software development. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.