ACTA IMEKO
ISSN: 2221-870X
December 2015, Volume 4, Number 4, 38-47

Key comparisons: the chance for discrepant results and some consequences

Franco Pavese

Torino, Italy

ABSTRACT
The paper reports an analysis of the results of complete sets of key comparisons available on the BIPM repository (KCDB), which shows that the majority of them show some discrepant results. The main reason of this fact is due to unknown or ill-estimated systematic effects, and indicates that the usual method of 'correcting' for known systematic effects is inefficient. The paper then discusses some of the possible reasons and an alternative approach, starting from a proposed revision of the meaning of systematic effect/error.

Section: RESEARCH PAPER

Keywords: key comparisons; BIPM KCDB; systematic effects; corrections

Citation: Franco Pavese, Key comparisons: the chance for discrepant results and some consequences, Acta IMEKO, vol. 4, no. 4, article 9, December 2015, identifier: IMEKO-ACTA-04 (2015)-04-09

Section Editor: Franco Pavese, Torino, Italy

Received April 7, 2015; **In final form** August 17, 2015; **Published** December 2015

Corresponding author: Franco Pavese, e-mail: frpavese@gmail.com

1. INTRODUCTION

The only means to verify the existence of significant systematic effects in the measurement results (in short results) of a laboratory is to perform comparisons of results of several (at least two) laboratories, in the assumption of lack of strong reasons for correlation of their results.

The Mutual Recognition Arrangement (MRA) [1] requires performing such an exercise, called "key comparison" (KC), as the main method to obtain statistical evidence of these effects. There are several categories of the KCs, the top rank being the CIPM KCs. Other types are the "regional", "bi-(multi-)lateral", "supplementary", etc.

A database (KCDB) of the results of these exercises is maintained at BIPM [2], which represents a unique and precious repository for understanding the state-of-the-art of each specific metrological field, and for the detection of significant systematic effects. In the MRA, the latter is called "systematic undetected differences" (SUD) between participants, requiring the application of a specific procedure in respect to the declaration of the Calibration and Measurement Capabilities (CMC), [1]. In general, the sets of KCDB results are also extremely valuable to assess the frequency of occurrence of significant systematic effects, or the lack of such occurrence.

This is "prior information" (not necessarily in the Bayesian sense) for the next exercises, in the statistical treatment of experimental data, namely of metrological ones. This issue should be taken in due account irrespective to the use of a frequentist, Bayesian or of any other approach, namely in the assessment of the quality of a measurement standard.

In this work an analysis of a large number of the KCs included in the KCDB is performed, as described in Sections 2 to 4. Then the paper discusses the reasons for inconsistent results in Sections 5 and 6, and proposes some remedies in Section 7, before the Conclusions are drawn. An Appendix acts as a digest useful to have at hand a summary of the current definitions of the main terms used in the paper.

2. A STATISTICS FROM THE BIPM KCDB

The analysis was performed on all the CIPM (master) KCs—some already including the relevant supplementary ones—included in the KCDB, with final results available at the date of November 15, 2012. They amounted to 339 of the total

Table 1. CIPM KCs from the KCDB.

KCDB	All OK	KCRV no-overlap	[x]	Pairs no-overlap	Marginal no-overlap	[y]
All CCs[a] (339)[b]	90	50	140 (41 %)	161 (41 %)	38	199 (59 %)
Auv (9)	4	4	8 (89 %)	0	1	1 (11 %)
Ccl (10)	2	1	3 (30 %)	6	1	7 (70 %)
Ccm (39)	10	6	16 (41 %)	20	3	23 (59 %)
Cct (7)	1	1	2 (29 %)	5	0	5 (71 %)
Em (45)	22	3	25 (53 %)	10	10	20 (47 %)
Pr (8)	0	3	3 (38 %)	5	0	5 (62 %)
Qm (162)	33	26	59 (36 %)	85	18	103 (64 %)
Ri (59)	18	6	24 (41 %)	30	5	35 (59 %)
[z]	27 %	73 %				

[a] Fields (including derived quantities): Auv: Acoustics, Ccl: Length, Ccm: Mass; Cct: Temperature; Em: Electromagnetics; Pr: ; Qm: Quantity of substance; Ri: Ionising radiation. In parentheses the total master KCs of that field.

[b] At November 15, 2012: grand total of all types 819.

[x] Proportion of KCs without participant-pairs non-overlap, but showing some non-overlap to the KCRV.

[y] Proportion of KCs with participant-pairs non-overlap (KCRV non-overlap not considered).

[z] Proportion of KCs with some non-overlap.

819 of all different types then included in the KCDB. They include all 8 areas in which the quantities are subdivided at the BIPM among the Consultative Committees (CC): Auv, Ccl, Ccm, Cct, Em, Pr, Qm, Ri.

The analysis concerns all the 'results', meaning all the specific comparison tables in the Final Result files (they may be one or dozens, depending on each KC)

Non-consistency is defined here as the non-overlap of the pairs of uncertainty intervals (provided for $k = 2$ in the KCDB). The number of occurrences of non-overlap of the results of each participant with respect to the Key Comparison Reference Value (KCRV, account being taken of its uncertainty) was recorded. The number of occurrences of non-overlap of the results of pairs of participants was separately recorded.

3. ANALYSIS OF THE KCDB RESULTS

Overall, the number of CIPM KCs with *no* inconsistencies is about *one fourth* only of the total. The number of CIPM KCs with *pair* inconsistencies is more than half (59 %). The anomalies are differently distributed for the KCRV non-overlap and for the pairs non-overlap. The statistics may be different for some metrological fields.

Table 1 reports a synthesis of the data collected. The corresponding typical overall *mixture* [3]–[5] (or pooled [6]) distributions of several CCT KCs is shown in Figure 1, together

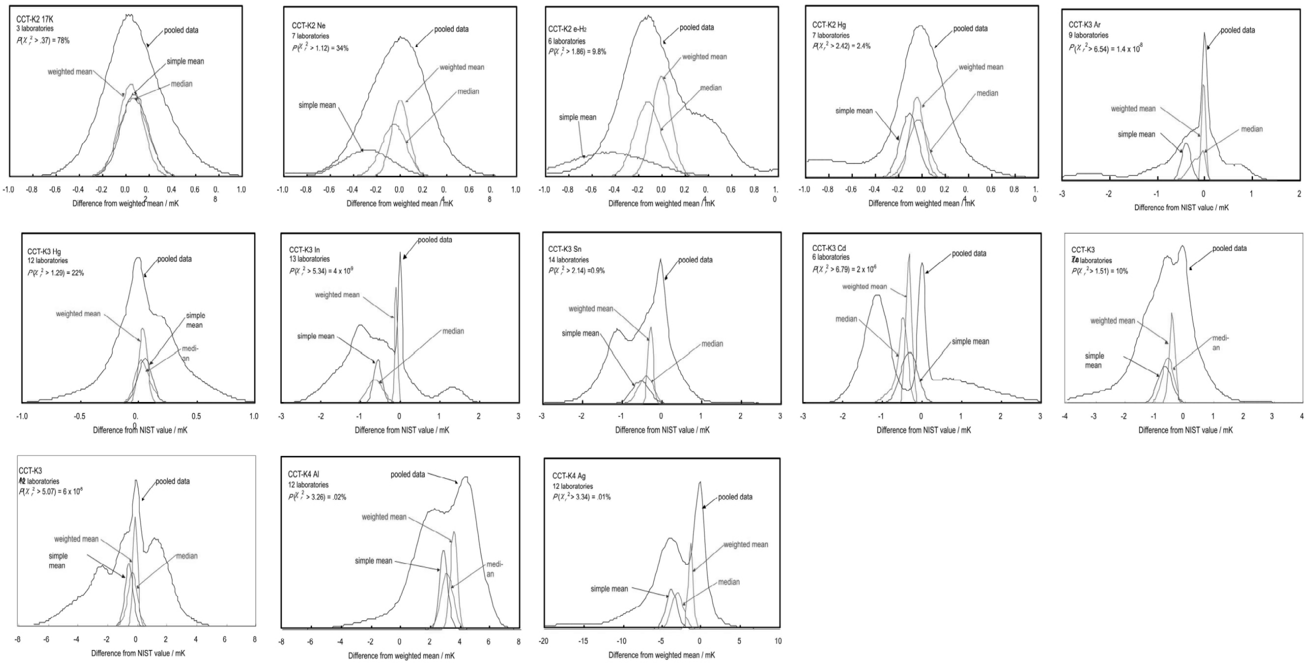

Figure 1. Mixture distributions for 13 fixed points in 3 CIPM CCT-KCs (temperature), from 7-20 local pdfs each (after [6]).

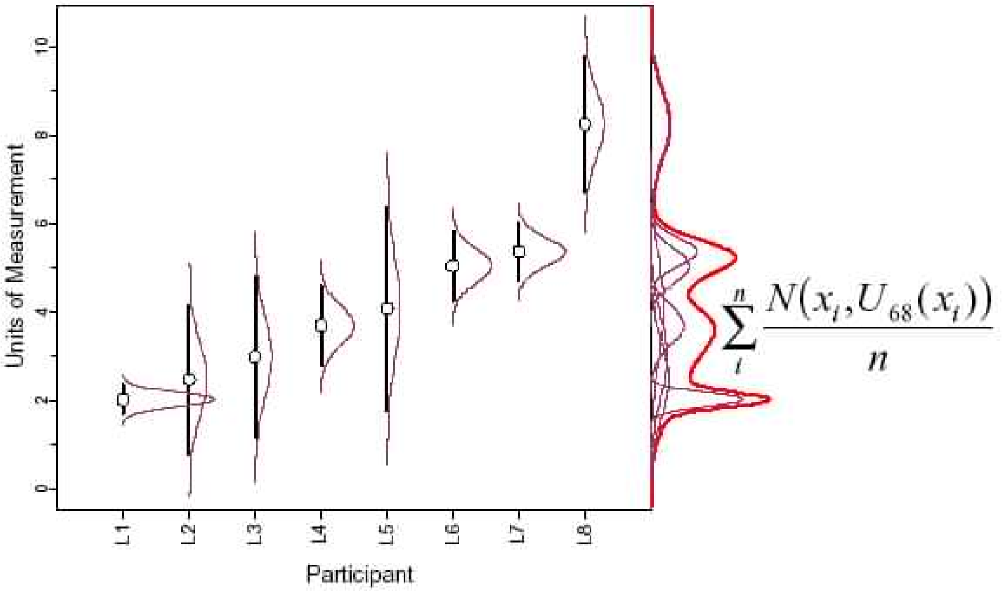

Figure 2. Results of a CCQM KC, showing on the right side the individual results and the mixture distribution (after [7]).

with the distributions of three common location parameters. Figure 2 shows a similar situation for a CCQM KC [7]. A more limited exercise of the same type was published in [8].

4. DISCUSSION OF THE RESULTS OF THE ANALYSIS

The literature data on ways to deal with this type of result inconsistencies is very wide; the reader can consult, e.g., [9]-[17]. Some of the proposed methods consist of criteria and procedures for modifying the results of the *inter*-comparison, or to find a subset of the overall results that is consistent and on which a representative reference value is then estimated.

However, while these may be reasonable solutions for generic *inter*-comparisons, that way-out is not allowed for the results of a MRA KC [1], which is a non-hierarchical *inter*-comparison designed mainly to obtain a *degree* of equivalence between the participants,[1] not their equivalence (called metrological compatibility in [18]). In fact, *all* originally-communicated results must be reported in the KC Final Report, and a participant can only possibly withdraw (but its participation remains recorded), or, should a so-called systematic unresolved difference (SUD) occur, can correct the uncertainty associated to its result (not the values), for the purpose of his application to include them into its CMCs.

In all instances, the validation of the laboratory results requires considering *inter*-comparisons as a mandatory step for the assessment of the quality of measurement standards.

On the other hand, the KC exercise in itself should be intended to check for the default, though top, accuracy of the measurement standards - i.e. it is a realisation under repeatability conditions, not an exceptional realisation - similarly to a "proficiency test" [19], but without being hierarchical.

An *inter*-comparison is a distinct step with respect to *intra*-laboratory information, consisting in the addition of *inter*-laboratory information gained by means of that exercise. It should therefore be included and discussed in the International Guidelines concerning the evaluation of the uncertainty of the results, like the GUM, [20] or in the relevant ISO standards. [21], [22]

5. SYSTEMATIC EFFECTS/ERRORS IN COMMON DATA-TREATMENT APPROACHES

Inconsistencies resulting in *inter*-comparisons may arise, in general, from:

- errors in the measurement of the *value* of influence quantities, responsible of systematic effects;
- under-evaluation of the associated *uncertainties*, often due misevaluated systematic effects;
- *omission* of considering some influence quantities, namely some of those responsible of systematic effects (model imperfection: epistemic problem).

The meaning of 'systematic effect' is assigned in both, the "uncertainty approach" (GUM [20]), in respect to the meaning of "input quantity" and "correction" and, the error approach (e.g., [18], [21]), where it is called "systematic error", in respect to the meaning of "bias" and "correction". It is necessary to understand first how these terms are defined in order to discuss their pros and cons.[2,3]

[1] It can be with respect to the KCRV, computed by using *all* the results, or as a difference between pairs of participants.

[2] The fact reported in the GUM, Note to clause 3.3.3 that "a 'random' component of uncertainty in one measurement may become a 'systematic' component of uncertainty in another measurement in which the result of the first measurement is used as an input datum" is not a justification for avoiding to introduce the systematic effects. See later on.

5.1. Uncertainty approach

According to the GUM approach, regarding the (mathematical) modelling of the measurement, clause (4.1) tells: "In most cases, a measurand Y is not measured directly, but is determined from N other quantities $X_1, X_2, ..., X_N$ through a functional relationship f:

$Y = f(X_1, X_2, ..., X_N)$ " (GUM, clause 4.1.1) (1)

and, "the input quantities $X_1, X_2, ..., X_N$ upon which the output quantity Y depends may themselves be viewed as measurands and may themselves depend on other quantities, including corrections and correction factors for systematic effects" … "The function f as it appears in this Guide is to be interpreted in this broader context, in particular as that function which contains every quantity, including all corrections and correction factors [*the systematic effects*] that can contribute a significant component of uncertainty to the measurement result" (GUM, clause 4.1.2).

However, after corrections are recognised in the GUM as part of the measurement model, in clause (3.2.3) it is also indicated: "It is assumed that, after correction, the expectation or expected value of the error arising from a systematic effect is zero": this clause mandates corrections deserving a preliminary step where the initial model is changed, as shown in the following simple example [25]:

Model after GUM 4.1.2: $Y = (X_1 + C_1) + (X_2 \cdot C_2)$, (2)

where X_1 and X_2 are two 'input quantities', C_1 and C_2 are a "correction" and a "correction factor" to X_1 and X_2, respectively;

Model after GUM 3.2.3: $Y \approx (X_1{}^* + C_1{}^*) + [X_2{}^* \cdot (1 + C_2{}^*)^2]$, (3)

where $X_1{}^* = X_1 + \mathrm{E}(C_1)$; $X_2{}^* = X_2 \cdot \mathrm{E}(C_2)$; $C_1{}^* = C_1 - \mathrm{E}(C_1)$; $C_2{}^* = C_2 - \mathrm{E}(C_2)$, where, in general, $\mathrm{E}(C_1) \neq 0$ and $\mathrm{E}(C_2) \neq 0$.

According to the GUM, the new quantities $C_1{}^*$ and $C_2{}^*$ have zero expectation: $\mathrm{E}(C_1{}^*) = 0$ and $\mathrm{E}(C_2{}^*) = 0$.

Expression (3) is the actual GUM model subject to the uncertainty analysis.

The corrections C_n (uncertain, according to GUM clause 3.3.1 "The result of a measurement after correction for recognized systematic effects is still only an estimate of the value of the measurand because of the uncertainty arising from random effects and from imperfect correction of the result for systematic effects.") [4] are "compensations", as said in VIM3 (2.53), "for an estimated systematic effect".

Actually, they are compensations for '*deviations*' of measured from a '*reference condition*' of specified quantities,[5] due to perturbations caused by other influence quantities producing off-set conditions, called systematic effects (the term 'bias' is not used in the GUM).[6] In short, the reference condition can be defined as the one where all the perturbations caused by other influence quantities are null (thus no corrections).

5.2. Error Approach

In the error approach, the '*deviations*' caused by the systematic effects are customarily called "measurement bias" (in short 'bias', see also Footnote 4), B_n, and corrections are taken as $C_n = - B_n$ (but see also Footnote 7 later on).

In the VIM3 (2.18) "measurement bias" is defined "*estimate* of a systematic measurement error" (emphasis added), where the "systematic measurement error" (in short 'systematic error') is defined in (2.17) "component of measurement error that in replicate measurements remains constant or varies in a predictable manner", and with Note 3 to (2.17) indicating "Systematic measurement error equals measurement error minus random measurement error".

In turn, the "measurement error" (in short error) is defined in VIM3 (2.16) "measured quantity value minus a reference quantity value" and "random measurement error" (in short 'random error') (2.19) "component of measurement error that in replicate measurements varies in an unpredictable manner".

Thus, the error is referred to each single observation of the set, while the distinction between systematic and random can arise only from "replicated measurements". Consequently, according to these definitions, the n-th variable representing the *population of the errors* $\varepsilon_{i,n}$, E_n, has the systematic error as expectation $\mathrm{E}(E_n)$, and the one that comes from the distribution of the errors as random error.

The bias, as defined in (2.18), can be an "estimate" of an error only if the difference between the "measured quantity value" and the "reference quantity value" cannot be known

[3] See the Appendix to this paper. Note that the GUM terminology is based on the VIM2 [36], obviously not on the subsequent VIM3.

[4] In GUM clause 3.2.3 a Note states: "The *uncertainty of a correction* applied to a measurement result to compensate for a *systematic effect* is not the *systematic error*, often termed *bias*, in the measurement result due to the effect as it is sometimes called. It is instead a measure of the uncertainty of the result due to *incomplete knowledge* of the required value of the correction. The error arising from imperfect compensation of a systematic effect cannot be exactly known. The terms 'error' and 'uncertainty' should be used properly and care taken to distinguish between them" (emphases added). It is true—see Section 5.2—but actually the knowledge is always incomplete.

[5] The GUM JCGM 104:2009 (GUM Supplement 4) [28], issued years after [20], admits in clause 3.11 that "Correction terms should be included in the model when the conditions of measurement are *not exactly as stipulated.* These terms correspond to systematic error values" (emphasis added). Thus a "stipulated" condition is introduced, looking consistent with the term used here, 'reference condition', but the need is confirmed that, "given an estimate of a correction term, the relevant quantity should be corrected by this estimate". Notice that this definition of "influence quantity" is not the VIM3 one (2.52) "quantity that, *in a direct measurement*, does not affect the quantity that is actually measured, but affects the relation between the indication and the measurement result" (emphasis added), but the GUM one, clause B.2.10 "quantity that is not the measurand but that affects the result of the measurement".

[6] By definition, once the influence quantities have been identified in each specific case, no other quantity outside the set of the input quantities plus the quantities responsible for systematic effects/errors can significantly influence the value of the measurand.

exactly. However, the first is the measured value. The second, being a "quantity value used as a basis for comparison with values of quantities of the same kind" (5.18), must be estimated only in the case (a) of the Note 2 to (5.18): "a reference quantity value can be a true quantity value of a measurand …", while in the case (b) "…is usually provided …", i.e. is an exact value—though it might have an associated uncertainty. Finally, bias is here relative to "a reference quantity value", and, as indicated in Note 2 to (2.17), "A correction *can* be applied to compensate for a known systematic measurement error" (emphasis added). As defined, the term bias looks not a variable since it is referred to each realisation, as the term error is.

In the ISO 3534:2006 (3.3.2) [21], "bias" is defined "difference between the expectation of a test result or measurement result and a true value". "Error of result" is defined in (3.4.4) "test result or measurement result minus the true value", and is made of two components: (3.4.6) the "random error of result" is the "component of the error of result which, in the course of a number of test results or measurement results, for the same characteristic or quantity, varies in an unpredictable manner" with "NOTE It is not possible to correct for random error"; (3.4.7) the "systematic error of result" is the "component of the error of result which, in the course of a number of test results or measurement results, for the same characteristic or quantity, remains constant or varies in a predictable manner".

Finally, "correction" is defined in (3.1.16) "action taken to eliminate a detected nonconformity" with "NOTE 1 A correction can be made in conjunction with corrective action" and "NOTE 2 A correction can be, for example, *reworked* or *regraded*", the definitions of the terms in italics being in ISO 9000:2005, 3.6.6. In turn, "nonconformity" is defined (3.1.11) "non-fulfilment of a requirement" and "corrective action" is defined (3.1.15) "action to eliminate the cause of a detected nonconformity or other undesirable situation".

Comparing the term "bias" (3.3.2) with the term "error" (3.4.4), the only difference is that *the first is the expectation of the second*, but in the second there is a single "true value" while in the first "a true value" is indicated. However, Note 1 to (3.3.2) states "Bias is the total systematic error as contrasted to random error": its meaning can be ambiguous. In fact, in (3.4.2) "measurement result" is defined "value of a quantity obtained by carrying out a specified measurement procedure", where "observed value" in (3.2.8) is "obtained value of a quantity or characteristic", which can be assimilated to "value" from NOTE 2 to (3.2.8) "Observed values may be combined to form a test result or measurement result".

Therefore, the measurement result is, in general, a set of observations, thus the term bias is not assigned to each single observation, being this consistent with the use of the term 'expectation' of the variable ('observed differences between test or measurement values and a true value'). On the contrary, Note 1 to (3.3.2) does not indicate the same concept, unless the systematic error is defined as the expectation of the error and the random error is defined as the variability of the error.

5.3. Meaning of bias due to systematic effects in this paper

There are two basic issues about the differences in the meaning of the term bias in ISO 3534:2006 that need be noted with respect to the VIM3 relevant set of definitions.

The first issue is that the term bias looks not to be assigned to a variable in the VIM3, while it is in the ISO 3534:2006. In addition, the latter specifies that it concerns the *total* effect of the systematic errors, so that the use of the plural 'biases' may be unnecessary.

The second issue is that bias in the ISO 3534:2006 is the deviation from "a true value", though Note 3 to (3.3.2) tells "In practice, the accepted reference value is substituted for the true value", being the latter unknown, while in the VIM3 it is the deviation form a "reference quantity value".

If the aim of the measurement is not to hunt for the true value, but to provide a value of the measurand according to the best estimated model of it and of the measurement system, the meaning of bias in the second issue is the same in both VIM3 and ISO 3534:2006 with respect to a reference value.

Concerning the first issue, it is more difficult to understand if the two definitions are equivalent, because of the ambiguity in the VIM3.

In this paper, the term "bias" is assumed to be represented by random variables, B_n, one for each of the relevant quantities involved in the model. See the Appendix to this paper.

In addition, the set of influence quantities, according to the model set in each specific "design of experiment", in the presence of bias is subdivided, as shown, e.g., in Section 5.1, into two sets: input quantities and quantities requiring a correction, the latter coincident with those responsible of systematic effects/errors. As said in Section 5.1, the 'deviations' are between the measured values and a 'reference condition' for each of the quantities responsible for a bias component. Non-zero bias means that the reference condition is evaluated not having been met (because a deviation is actually observed); or, that an unexpected deviation from a 'standard condition' occurs, so that the measurement results are systematically 'off centre' in the variations of their values.

Finally, it should be clear that the term "systematic" should not be taken as 'fixed': at best one can use the expected value of a distribution of its values.

6. ANALYSIS OF THE REASONS OF FAILURE OF THE PRIOR-CORRECTION APPROACH

The historical practice of requiring a prior correction for the systematic effects/errors has been shown to fail in the majority of cases of the CIPM KC results—also representative of other similar exercises. That practice does not ensure sufficient confidence in taking effectively into account the systematic effects.

The reason of failure looks residing in an insufficient understanding of two basic features of a reliable process of evaluation of measurement data:

(a) the overall, single-step nature of the analysis required by the 'design of experiment', forming the *within*-laboratory knowledge framework;

(b) the intrinsic need to consider the *inter*-comparison of independent results as an integral final step in the *between*-laboratory subsequent framework of the process, before a statement about result consistency (metrological compatibility [18]) can be assessed with sufficient confidence/belief.

The feature (a) will not be discussed here in details and is deferred to a subsequent study [29]. It seems sufficient to recall here that the process of identifying the influence quantities is not a two-step one, the first concerning what in the GUM are called the "input quantities" and the second concerning the quantities responsible for the systematic effects calling for corrections. In addition, it is a fact that no measured or evaluated value of a quantity, including corrections, can be exempt from ignorance as to the actual location of the true value, or from uncertainty arising from the dispersion of the measured data, nor can any evaluation be considered exact, except by convention, irrespective to the process bringing to it (either Type A or Type B).

Therefore, in a typical model no difference in principle should exist between the influence quantities called "input quantities" X_i and those responsible of "systematic effects", called biases B_n in the error approach, and said to require prior correction in both approaches. [23]-[27] Each of them can in fact be expressed as follows

$$\text{(a) } X_i = \mathrm{E}(X_i) + X_i^*, \qquad \text{(b) } B_n = \mathrm{E}(B_n) + B_n^*, \qquad (4)$$

perfectly equivalent with each other, where the starred quantities have zero mean and take into account the variability of the data—the normally non-zero mean (the 'quantity value') has been 'extracted' from the non-starred quantities. In (a) $\mathrm{E}(X_i)$ is the estimate of the quantity value, in (b) $\mathrm{E}(B_n)$ is the estimate of the systematic component of the effect/error, i.e. of the difference between the value of the measured quantity and a reference value; in (a) X_i^* is what is normally indicated as E_i, the random error, in (b) B_n^* is what is indicated in the standards and guidelines as the random component of the effect/error.

The bias indicates the deviations of the measured values from a set of reference values forming what was called hereinbefore a 'reference condition'. For example, for a single additive bias B_i, $X_i = {}^{\ominus}X_i + B_i$ (the symbol of 'standard state' $^{\ominus}$ is borrowed here, for analogy, from physical chemistry to indicate the reference condition for which $\mathrm{E}(B_i)$ =: 0), and ${}^{\ominus}X_i + B_i = {}^{\ominus}X_i - C_i$. [7]

7. A CONSEQUENCE

The resulting model, rather than (1), should be:

$$Y = f(W_1, W_2, \ldots, W_m, \ldots, W_M), \qquad (5)$$

where W_m are either the $X_1, \ldots, X_I$ and the $C_1, \ldots, C_N$, with $i = 1\ldots I$, $n = 1\ldots N$, $M = I + N$: the W_m include the 'correction' terms, carrying in also an uncertainty component.[8]

The only characteristics of bias is functional (e.g., in a cause-effect diagram): it applies to an 'input quantity', or to another bias term. The practical difference between correction C_n and input quantity X_i basically is: (i) for an "input quantity", X_i, the measurement results always involve one or more Type A components; (ii) for a correction term, C_n, the evaluation may involve only Type B components.

For any random variable, the expectation of the measured or estimated dispersion of the data does not allow, *within* each set of measurements, to estimate with sufficient confidence that no bias exists due to missed or ill-evaluated systematic effects. Therefore, a model like (5) only eliminates the confusion that can arise from an incorrect use of the corrections, but does not alleviate the problem of assessing the consistency of independent measurements.

That assessment, which can be called the *between*-laboratory traceability, is a basic need in metrology, as no isolated measurement can ensure it—remaining in practice useless. According to the above feature (b), only *inter*-comparisons can increase the confidence in the correctness of the assessment. [30]–[33] On the contrary, as said at the beginning of Section 4, most of the published effort has been devoted to find methods to deal with inconsistent data, giving for granted the separate treatment of the systematic effects, though their definitions were not univocal, as shown in Section 5.

8. CONCLUSIONS

The analysis of the KCDB indicated that the method of the prior correction of systematic effects fails utterly in the majority of cases.[9] The KCDB prior information—that the KCs outcomes are most likely to reveal non-consistent data—cannot be omitted from the measurement model, nor one can assume with sufficient confidence that "after correction, the expectation or expected value of the error arising from a systematic effect is zero".

A different more robust way should be applied to the treatment of the measurement results, irrespective to the approach used for the treatment (frequentist, Bayesian or else; error or uncertainty).

All 'corrections' should be considered what they actually are: imprecise compensations of bias, a type of 'input quantity' to be included in the model as any other influence quantity and treated as such. The (true) value of a 'correction' remains as unknowable as that of any 'input quantity', and as that of the measurand itself. Incidentally, the estimated value of a correction can be set to zero[10] (randomised) to mean *ignorance* about its value (and sign), but it always carries an uncertainty taking also this fact in due consideration. The rational for that is the fact that the bias originates from systematic effects (at least the non-epistemic ones) and should be treated as a random variable of the same nature of those of the variables modelling the 'input quantities', as shown in (4). A separate step for applying some corrections (meaning by using $\mathrm{E}(C_n)$) is not forbidden, but in principle is unnecessary and prone to possible confusion.

The common characterisation of the systematic effects in sentences like, e.g., "the 'systematic' sources are such parameters associated with calibration standards, reference data,

[7] Notice that, in general, $\mathrm{E}(C_i)$ is intended for 'correction'.

[8] Whether should the model include the corrections or the systematic effects is something worth to ponder about. It depends on the purpose of the model. The GUM one looks not to be the 'measurement model' in the sense of a model constructed *prior* the measurements are performed—which should use the (physical) influence quantities (X_i and B_n in the notation of this paper)—but the *posterior* model of the measurement outcomes—using the measured (or evaluated) influence quantities M_i, and C_n. [29]

[9] In [35] a systematic effect is said to "constitute an element for which treatment conventional statistics fails utterly", where 'conventional' stands for 'non-Bayesian'. Author's opinion is that this paper has shown that the statement might not be true.

[10] This fact indicates that, differently from the 'input quantities', the values of the influence quantities responsible for the systematic effects are not used, only their differences with respect to the reference state.

bias corrections, etc., whose effects are usually common and not altered under repeatability conditions" [34] implies that no uncertainty is associated to the effects. This is not true: also Type B evaluations (in GUM language) are based, directly or indirectly, explicitly or implicitly, on experimental data or on subjective judgement, both uncertain. Each of the estimates of the influence quantities appearing in the model is affected by their own uncertainty. [11]

An alternative way has been suggested in this paper, the treatment of an overall (single) model including as a single one the two sets of influence quantities in which the overall set is traditionally subdivided: the input quantities and the bias quantities (the latter traditionally suggested to require prior correction, a practice that should be abandoned). The model can be broken into sub-models only for a matter of convenience in computation, but only in very simple cases without basic implications for the overall treatment. A separate step for applying some corrections (meaning by using E(C_n)) is not forbidden, but is in principle *unnecessary*.

As a final step, the result assessment process must always be integrated with *inter*-comparisons. That need does not concern only metrological studies, but also studies in other fields, like testing (e.g., assessment of reference materials), see [37].

Appendix

Meaning of some critical terms used in this paper with respect to their definitions in the VIM3 [2], GUM [1] and ISO 3534-2:2006 [9]

There are persisting difficulties in terminology arising from the fact that the same concept can be found defined differently in different International Guidelines and Standards. This fact is partly due because the current versions of these documents, some regulatory some informative, date back to different times. For the ones taken in consideration here, the VIM3 [2] basically dates 2008, the GUM [1] dates 1998 (*so the terms are defined in it according to the* VIM2 [3], issued in 1998), and the ISO 3534-2 dates 2006.

As a consequence, not necessarily all the relevant definitions are consistent with each other.

It is not the aim of this Appendix to discuss the issue in full, but only to compare in brief the above documents with the meaning that the author have chosen to assign in this paper to certain critical terms. However, no term is used in this paper with a meaning not conforming at least one of the existing Guidelines or Standards.

The Appendix is not in the alphabetical order of the terms because the illustration may require interrelating them logically.

Measurement method

According to the VIM3 term 2.5, one can have "direct measurement methods, and indirect measurement methods". This paper only refers to indirect methods. Actually, in the vast majority of cases and for sufficiently high accuracy a direct method is only an ideal approximation, since, due to the **corrections**,[12] the number of **influence quantities** is higher than that strictly required for the direct method.

Influence quantity

In this paper, we include in this term all quantities whose effect has a significant influence on the numerical value of the "measurement result" (VIM3 term 2.9), according to its "target uncertainty level" (VIM3 term 2.34), and irrespective to the fact that the numerical values are provided by measurement, or are obtained with methods requiring Type B uncertainty evaluations.

For the reasoning in this paper, we do not consider appropriate the VIM3 term 2.52 definition "quantity that, *in a direct measurement*, does not affect the quantity that is actually measured, but affects the relation between the indication and the **measurement result**" (emphasis added, see **measurement method**). The meaning in this paper is basically the one in the GUM term B.2.10 "quantity that is not the measurand but that affects the result of the measurement", taken from the VIM2. Thus, this set of quantities is broader than that of the **input quantities** (see section 2).

The ISO 3534-2:2006 does not define this term.

Measurement result

This paper uses the VIM3 2.9 definition "set of quantity values being attributed to a measurand together with any other available relevant information".

The ISO 3534-2:2006 term 3.4.2 definition, "value of a quantity obtained by carrying out a specified measurement procedure", is different, and the GUM considers the measurement result basically as single valued (e.g., clause 4.1.5).

Measurement model

In this paper, it is the model including all the influence quantities. The VIM3 term 2.48 definition looks similar, but does not have exactly the same meaning: "mathematical relation among all quantities known to be involved in a measurement". Clearly this type of model depicts indirect measurements.

The GUM clause 3.1.6 tells: "The mathematical model of the measurement that transforms the set of repeated observations into the measurement result is of critical importance because, in addition to the observations, it generally includes various influence quantities that are inexactly known. This lack of knowledge contributes to the uncertainty of the measurement result, as do the variations of the repeated observations and any uncertainty associated with the mathematical model itself". Clauses 4.1.1–4.1.3 depict the same model of the VIM3, but with different definitions for its components (see **input quantity**, **correction**).

[11] In GUM Appendix E, the clause E.3.5 states "In the traditional terminology, the third term on the right-hand side of Equation (E.6) is called a 'random' contribution to the estimated variance $u^2(z)$ because it normally decreases as the number of observations n increases, while the first two terms are called 'systematic' contributions because they do not depend on n". It is correct for n observations of another quantity, but an increasing number of observations, concurring to estimate the expectation of a quantity responsible for a systematic effect, concurs in decreasing its uncertainty.

[12] As in VIM3, the bold type is used here for other terms illustrated in the Appendix.

Measurement function

In this paper, it is the functional relationship implementing the measurement model.

In the VIM3 term 2.49, it is a "function of quantities, the value of which, when calculated using known quantity values for the **input quantities** in a **measurement model**, is a measured quantity value of the output quantity in the measurement model". "NOTE 1 If a measurement model $h(Y, X_1, \ldots, X_n) = 0$ can explicitly be written as $Y = f(X_1, \ldots, X_n)$, where Y is the output quantity in the measurement model, the function f is the measurement function. More generally, f may symbolize an algorithm, yielding for input quantity values $x_1, \ldots, x_n$ a corresponding unique output quantity value $y = f(x_1, \ldots, x_n)$".

It is basically the same in the GUM.

Input quantity

This paper uses the term in the meaning of the GUM, as expressed in clause 4.1.2: "the *input quantities* $X_1, X_2, \ldots, X_N$ upon which the output quantity Y depends may themselves be viewed as measurands and may themselves depend on other quantities, including **corrections** and **correction factors** for **systematic effects**, thereby leading to a complicated functional relationship f that may never be written down explicitly. Further, f may be determined experimentally or exist only as an algorithm that must be evaluated numerically. The function f as it appears in this Guide is to be interpreted in this broader context, in particular as that function which contains every quantity, including all corrections and correction factors that can contribute a significant component of uncertainty to the measurement result. Thus, if data indicate that f does not model the measurement to the degree imposed by the required accuracy of the **measurement result**, additional input quantities must be included in f to eliminate the inadequacy. This may require introducing an input quantity to reflect incomplete knowledge of a phenomenon that affects the measurand" (emphasis added).

The VIM3 term 2.50 definition is "quantity that must be measured, or a quantity, the value of which can be otherwise obtained, in order to calculate a measured quantity value of a measurand", with "NOTE 2 Indications, corrections and influence quantities can be input quantities in a measurement model". The different meaning here of **influence quantity** is evident, so is the restricted meaning of "indication" (term 4.1). See the term **correction** for its relationship to input quantity.

The ISO 3534-2:2006 does not define this term.

Correction, correction factor

Correction is the key term analysed and discussed in this paper, where it is treated as a random variable, according to the GUM clause 4.1.2 "The input quantities $X_1, X_2, \ldots, X_N$ upon which the output quantity Y depends may themselves be viewed as measurands and may themselves depend on other quantities, including **corrections** and **correction factors** for **systematic effects**". Thus in this paper the quantities responsible for corrections are not included among the **input quantities**, while they are among the **influence quantities** (but, e.g., GUM clause 4.1.3 "…corrections for influence quantities…") and expressing **systematic effects** (e.g., GUM clause E.1.1 "correction for a systematic effect"). Apparently "correction" is only for an additive correction term and "correction factor" only for a scale-factor correction term.

In the VIM3 term 2.53 is defined as "*compensation* for an estimated **systematic effect**. NOTE 1 See GUM:1995, 3.2.3, for an explanation of 'systematic effect'. NOTE 2 The compensation can take different forms, such as an addend or a factor, or can be deduced from a table" (emphasis added).

In the VIM3 term 2.50 input quantity saying "NOTE 2 Indications, corrections and influence quantities can be input quantities in a measurement model", does not exactly match the GUM definition. In the VIM3 the term **correction** only comes in the following terms: 2.3 Measurand, 2.17 Systematic measurement error, 2.26 Measurement uncertainty, 2.39 Calibration, 2.50 Input quantity, 2.53 Correction and 5.10 Intrinsic measurement standard.

The ISO 3534-2:2006 term 3.1.16 definition is "action taken to eliminate a detected nonconformity", where the term "nonconformity" (3.1.11) definition is "non-fulfilment of a requirement".

None of the above documents indicates what exactly the "action" or "compensation" means.

As to uncertainty of the correction, the GUM clause 3.2.3 NOTE states: "The uncertainty of a **correction** applied to a **measurement result** to *compensate* for a **systematic effect** is not the ***systematic error**, often termed **bias***, in the measurement result due to the effect as it is sometimes called. It is instead a measure of the uncertainty of the result due to incomplete knowledge of the required value of the correction. The error arising from imperfect compensation of a systematic effect cannot be exactly known." (emphasis added).

As to the accuracy of the correction, the GUM clause 3.2.3 states: "it is assumed that, after correction, the expectation or expected value of the error arising from a systematic effect is zero". The VIM term 2.17 Note 2 states "**Systematic measurement error**, and its causes, can be known or unknown. A **correction** *can* be applied to compensate for a known systematic measurement error." (emphasis added), apparently different from GUM.

Measurement error

In this paper this term has the meaning defined in the ISO 3534-2:2006 term 3.4.4 "test result or measurement result minus the true value".

The VIM3 term 2.16 definition is "measured quantity value minus a reference quantity value", different from the ISO 3534-2:2006 term 3.4.4., irrespective to the fact that the latter also tells "NOTE 1 In practice, the accepted reference value is substituted for the true value".

GUM, not adopting the term, tells in clause 3.2.1. "In general, a measurement has imperfections that give rise to an error in the measurement result. *Traditionally*, an error is viewed as having two components, namely, a random component and a systematic component. NOTE Error is an idealized concept and errors cannot be known exactly" (emphasis added).

Systematic effect, systematic error, bias

In this paper, the term **systematic effect** is used with the same meaning it has in the GUM clause E.3.4 (after equation E.4): "Here α is a constant "systematic" offset or shift common to each observation, and β is a common scale factor. The offset and the scale factor, although fixed during the course of the observations, are assumed to be characterized by a priori probability distributions, with α and β the best estimates of the expectations of these distributions". Thus these effects are random variables. Systematic errors are caused by systematic effects. Also relevant here is the GUM clause E.1 in full.

In this paper, the term **bias** is used with the same meaning it has in the ISO 3534-2:2006 term 3.3.2 “difference between the expectation of a test result or measurement result and a true value. NOTE 1 Bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the true value is reflected by a larger bias value. NOTE 3 In practice, the accepted reference value is substituted for the true value”. However, Note 3 is applied.

The VIM3 term 2.18 definition of measurement bias is “estimate of a systematic measurement error” and term 2.17 definition of the latter is “component of **measurement error** that in replicate measurements remains constant or varies in a predictable manner. NOTE 1 A reference quantity value for a systematic measurement error is a true quantity value, or a measured quantity value of a measurement standard of negligible measurement uncertainty, or a conventional quantity value. NOTE 2 Systematic measurement error, and its causes, can be known or unknown. A **correction** can be applied to compensate for a known systematic measurement error. NOTE 3 Systematic measurement error equals measurement error minus random measurement error”. See the text above for a discussion about the meaning of clause 2.18.

In the GUM clause 3.2.3 NOTE the specification “…**systematic error**, often termed **bias** …” indicates that the term bias is not recommended, being relative to an error, and is not used.

Standard state, reference condition, nominal value

This paper introduces the concept of standard state in Section 6 and of reference condition in subsection 5.1, as a condition where known significant systematic effects are absent. (not simply null) Bias, as an out-of-reference condition, is introduced in subsection 5.2.

The GUM clause 5.1.5 introduces a *nominal value*: “If Equation (1) [*the* ***measurement model***] for the measurand Y is expanded about nominal values $X_{i,0}$ of the input quantities X_i, then, to first order (which is usually an adequate approximation), $Y = Y_0 + c_1\delta_1 + c_2\delta_2 + \ldots + c_N\delta_N$, where $Y_0 = f(X_{1,0}, X_{2,0}, \ldots, X_{N,0})$, $c_i = (\partial f/\partial X_i)$ evaluated at $X_i = X_{i,0}$, and $\delta_i = X_i - X_{i,0}$. Thus, for the purposes of an analysis of uncertainty, a measurand is usually approximated by a linear function of its variables by transforming its input quantities from X_i to δ”.

However, except for the above operational illustration, in the GUM the meaning of nominal value is not defined. The δ_i are not labelled **corrections**. In the VIM3 term 4.6 **nominal quantity value** definition is “rounded or approximate value of a characterizing quantity of a measuring instrument or measuring system that provides guidance for its appropriate use” and looks having a meaning different from the previous.

REFERENCES

[1] CIPM, Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes (MRA). (1999) Bureau International des Poids et Mesures, Sèvres. http://www.bipm.org

[2] http://kcdb.bipm.org

[3] P. Ciarlini, G. Regoliosi and F. Pavese: “A bootstrap algorithm for mixture models and interval data in inter-comparisons”, Proceedings A4A4 Conference, Huddersfield, UK, 2002, pp. 138-145.

[4] P. Ciarlini, M.G. Cox, F. Pavese and G. Regoliosi: “The use of a mixture of probability distributions in temperature interlaboratory comparisons”, Metrologia, 41 (2004) pp. 116-121.

[5] B. Toman, Bayesian Approach to Assessing Uncertainty and Calculating a Reference Value in Key Comparison Experiments, J. Res. Natl. Inst. Stand. Technol. 110 (2005) pp. 605-612

[6] A.G. Steele, K.D. Hill, R.J. Douglas, Data Pooling and key comparison reference values, Metrologia 39 (2002) pp. 269–277

[7] D.L. Duewer, A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers, Accred. Qual. Assur. 13 (2008) pp. 193–216

[8] A. Hornikova, W.F. Guthrie, A Survey of Design, Analysis, and Reporting of Results in Key Comparisons, XVIII IMEKO World Congress, September, 17–22, 2006, Rio de Janeiro, Brazil

[9] K. Beissner, On a measure of consistency in comparison measurements: II. Using effective degrees of freedom, Metrologia 40 (2003) pp. 31-35

[10] M.G. Cox, A.B. Forbes, J.L. Flowers, P.M. Harris, Least squares adjustment in the the presence of discrepant data, in Advanced Mathematical and Computational Tools in Metrology VI” (P. Ciarlini, M.G. Cox, F.Pavese, G.B. Rossi, Eds."), vol.6, Series on Advances in Mathematics for Applied Sciences vol.66, World Scientific, Singapore, 2004, pp. 37-51. ISBN: 981-238-904-0

[11] A.G. Steele, R.J. Douglas, Consistency measures for peer-to-peer comparisons, XVIII IMEKO WORLD CONGRESS, September, 17 – 22, 2006, Rio de Janeiro, Brazil

[12] M.G. Cox, The evaluation of key comparison data: determining the largest consistent subset, Metrologia 44 (2007) pp. 187-200

[13] A.G. Chunovkina, C. Elster, I. Lira, W. Wöger, Analysis of key comparison data and laboratory biases, Metrologia 45 (2008) pp. 211-216

[14] R.N. Kacker, A.B. Forbes, R. Kessel K.-D. Sommer, Bayesian posterior predictive p-value of statistical consistency in interlaboratory evaluations, Metrologia 45 (2008) pp. 512–523

[15] A.G. Chunovkina, N. Zviagin, N. Burmistrova, Evaluation of inconsistent data of key comparisons of measurement standards, Acta IMEKO 2012 TC21

[16] C. Elster, B. Toman, Analysis of key comparison data: critical assessment of elements of current practice with suggested improvements, Metrologia 50 (2013) pp. 549–555

[17] R. Willink, Models for the treatment of apparently inconsistent data, in "Advanced Mathematical and Computational Tools in Metrology and Testing X” (F.Pavese, W. Bremser, A.G. Chunovkina, N. Fischer, A.B. Forbes, Eds.), vol.10, Series on Advances in Mathematics for Applied Sciences vol 86, World Scientific, Singapore, 2015, pp. 78-89

[18] International vocabulary of metrology – Basic and general concepts and associated terms 3rd edition (VIM3) 2008 (electronic version JCGM 200:2012 from the BIPM/JCGM is available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf)

[19] ISO 13528:2005 and ISO 17034:2010, International Organization for Standardization (Geneva, Switzerland) 2005 and 2010

[20] Guide to the Expression of Uncertainty in Measurement (GUM) 2nd edn 1995 (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML) ISBN 92-67-10188-9 (electronic version JCGM 100:2008 from the BIPM/JCGM is available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf)

[21] ISO 3534-2:2006, Statistics—Vocabulary and symbol —Part 2: Applied statistics, International Organization for Standardization (Geneva, Switzerland) 2006

[22] ISO 5725, 1994 2nd edn Accuracy (trueness and precision) of measurement methods and results, International Organization for Standardization, Geneva, Switzerland.

[23] F. Pavese, About the treatment of systematic effects in metrology”, Measurement 42 (2009) pp. 1459-1462

[24] F. Pavese, On the difference of meanings of “zero correction”: zero value vs no correction, and of the associated uncertainties, in “Advanced Mathematical and Computational Tools in

Metrology and Testing IX" (F.Pavese, M. Bär, J.-R. Filtz, A. B. Forbes, L. Pendrill, K. Shirono Eds.), vol. 9, Series on Advances in Mathematics for Applied Sciences vol. 84 World Scientific, Singapore, 2012 pp. 297-309

[25] F. Pavese, Why should correction values be better known than the measurand true value?, Journal of Physics: 459 (2013) 012036 doi:10.1088/1742-6596/459/1/012036

[26] F. Pavese, Why the distinction between corrections and input quantities, St. Petersburg Seminar 2012

[27] F. Pavese, Corrections and input quantities in GUM-compliant models, XX IMEKO World Congress, Metrology for Green Growth, September 9−14, 2012, Busan, Republic of Korea

[28] Evaluation of measurement data — An introduction to the "Guide to the expression of uncertainty in measurement" and related documents, 2009 (electronic version JCGM 104:2009 from the BIPM/JCGM is available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_104_2009_E.pdf)

[29] F. Pavese, A note on the classification in random errors and systematic errors, *in preparation.*

[30] F. Pavese, On some consequences of the different nature of within- and between-laboratory data, Metrologia 46 (2009) pp. L29-L32

[31] F. Pavese, Dependence of the treatment of systematic error in interlaboratory comparisons on different classes of standards, ACQUAL 15 (2010) pp. 305-315

[32] F. Pavese, On hierarchical vs. non-hierarchical comparisons in metrology and testing, Int. J. Metrol. Qual. Eng. 1 (2010) pp. 7–10

[33] F. Pavese, Mathematical and statistical tools in metrological measurement, 2013, Chapter in Physical Methods, Instruments and Measurements, [Ed. UNESCO-EOLSS Joint Committee], in Encyclopœdia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK, http://www.eolss.net

[34] In Jung Kim, Byungjoo Kim, Euijin Hwang, An Approach for the Uncertainty Evaluation of the Overall Result from Replications of Measurement: Separately Combining Individual Uncertainty Components According to their 'systematic' and 'random' Effects, Bull. Korean Chem. Soc. 35 (2014) pp. 1057–1060

[35] I. Lira, W. Wöger, Comparison between the conventional and Bayesian approaches to evaluate measurement data, Metrologia 43 (2006) pp. S249–S259

[36] International vocabulary of basic and general terms in metrology (VIM2), second edition, 1993, International Organization for Standardization (Geneva, Switzerland)

[37] D.L. Duewer, H. Gasca-Aragon, K.A. Lippa, B. Toman, Experimental design and data evaluation considerations for comparisons of reference materials, Accred. Qual. Assur. 17 (2012) pp. 567–588.