ACTA IMEKO
ISSN: 2221-870X
June 2018, Volume 7, Number 2, 73-83

Data inter-comparisons in the context of the knowledge-gaining process: an overview

Franco Pavese[1], Abderafi Charki[2]

[1] *Torino, 10139, Italy*
[2] *University of Angers, LARIS - 62 Avenue Notre Dame du Lac 49000 Angers, France*

ABSTRACT
This paper deals with the principle of data inter-comparisons, the object of which is to increase knowledge continuously with respect to time. Although the principle is as such nothing new to metrology and testing laboratories, which carry out experimental measurements, a degree of clarification is nonetheless called for in view of the numerous questions that arise concerning ways of implementing and utilizing it to improve knowledge capitalization.

The acquisition of knowledge relative to any measurand involves a series of steps: studying the state of knowledge of the measurand, choosing a working method (typically, by establishing a design of experiment), obtaining the measurements, and analyzing them. Following this, an action plan is established in order to reduce (or if possible avoid) weaknesses or over sensitivity.

Comparisons are already conducted using various approaches within a laboratory. It is therefore important to understand that to assess the accuracy of a method and validate it, it is necessary to compare the results obtained by several laboratories for a given method and measurand with the correct type of inter-comparison. It is this comparison between several laboratories that, when using different methods, produces the most up to date knowledge with the highest confidence level. This paper goes over the steps that allow developing knowledge, presenting the aims and characteristics of the various inter-laboratory comparison methods, notably referring to the tools established by documents such as the BIPM MRA (the Mutual Recognition Arrangement), the ISO 5725 and the ISO 13528.

Section: RESEARCH PAPER

Keywords: inter-comparisons; infra-comparisons, proficiency testing; accuracy; reliability; uncertainty; validation; competency; knowledge

Citation: Franco Pavese, Abderafi Charki, Data inter-comparisons in the context of the knowledge-gaining process: an overview, Acta IMEKO, vol. 7, no. 2, article 13, June 2018, identifier: IMEKO-ACTA-07 (2018)-02-13

Editor: Dušan Agrež, University of LJUBLJANA, Slovenia

Received January 16, 2018; **In final form** May 29, 2018; **Published** June 2018

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

1. INTRODUCTION

The measurand is a concept that is shared in the scientific and technical communities [1]–[4]. Any given measurand is supposed to be the object of replicated measurements that must be comparable with each other. In other words, it is, or should be, recognized as a quantity having a current recognizable meaning for the specific community. In the jargon of science philosophers, this means that it should be projected onto a “social framework” [5]. In the scientific experimental context, this means that the measurand model must be of the “prescriptive” type, meaning “giving directions or injunctions” [6], which does not always denote the “physical” model.

The confidence that can be ascribed to a measurement process is normally based on an estimation of uncertainty [7], [8]. However, uncertainty can often be difficult to assess properly in certain complex situations for which unique physical models do not a priori exist. And yet laboratories, customers and/or instructing parties need to have accurate information about the confidence level associated with measured data in order to make important decisions that are justified, technically and scientifically, about the validity of a method or the conformity of a procedure, to authorize the sale of a product, and endorse and maintain an individual’s competence and skills.

The core concept of measurement needs to be inseparably complemented by the need to ensure reliability of results (qualitative and quantitative) and, whenever possible, measurement traceability. That need obviously also implies a sound understanding of the concept of calibration and of comparison. Calibration [9] is an operation that, under specified conditions and as a first step, establishes a relationship between the quantity values, associated with measurement uncertainties provided by the measurement standards, and the corresponding indications with their associated measurement uncertainties. As a second step, this information is used to establish a relationship for obtaining a measurement result from an indication.

The accreditation standard for testing and calibration laboratories is ISO/CEI 17025 [10], which contains the requirements relating to management aspects and the technical requirements necessary to ensure the accuracy and reliability of results obtained by a laboratory.

Many factors can influence the accuracy and reliability of testing and/or calibration carried out by a laboratory [10]. Some of these influence factors may include the following:

- Level of competence of personnel
- Equipment and ambient conditions
- Handling and storage of objects for test and calibration
- Measurement traceability
- Sampling
- Collection
- Testing and calibration methods (whether developed or adopted, standardized or non-standardized)
- Means to be used to ensure the quality of testing and calibration results:
    - ✓ Regular use of certified reference materials and/or internal quality control with use of secondary reference materials
    - ✓ Participation in comparison programs between laboratories, or proficiency tests
    - ✓ Tests or calibration repeated using identical or different methods
    - ✓ Renewed tests or renewed calibration of stored objects
    - ✓ Correlation of results for different characteristics of an object.

It is important to remember that a quantity value [9] is the basis for the subsequent comparison of values of quantities of the same kind. In many situations, a known (conventional) quantity value can be used for a true quantity value of the measurand, in lieu of the true value that is always unknown. A conventional (reference) quantity value with associated uncertainty is usually assigned, for example, to a certified reference material, or for the characteristics of a device, e.g., a stabilized laser, or in the case of a reference measurement procedure or, often, for a comparison of measurement standards.

The concept of comparison has become widely used in many sectors, and many new proficiency test (PT) schemes are currently launched every year worldwide [11], [12]. A detailed study of comparison testing has been conducted [12]. The International Organization for Standardization (ISO) has published a reference document on general requirements for proficiency testing [13] and a standard on statistical methods for use in proficiency testing [3]. The International Laboratory Accreditation Corporation (ILAC) has also published a document on requirements for proficiency testing [14].

In the testing field, the questions most frequently asked about the subject of comparisons are the following. Does the possibility of a comparison in my case exist? With whom can I or should I compare? What should I do if no comparison solution has been organized? Who might be able to set up the comparison? Is an inter- or intra-laboratory comparison sufficient in my case? Is there a recognized reference? Is there a reference laboratory? What should be done if no reference laboratory exists? Who can prepare the samples to be tested? How can the stability and homogeneity of the reference material and/or samples sent to each laboratory be ensured? Can I compare with a single other organization? How many laboratories need to be involved for a comparison to be reliable? Do all the laboratories taking part in a comparison need to have ISO/IEC 17025 Standard [10] accreditation? Does the organizer of the inter-laboratory comparisons have the necessary competences? Should the organizer be independent?

The principle of comparison has long since been used and considered effective by national metrology institutes (NMI), notably, since 1999, in the form of Key Comparisons [1] organized under the auspices of the International Bureau of Weights and Measures [1]. Key comparisons of national standards are carried out to establish the *degrees of equivalence* of measurement results obtained by the NMIs [4], [15], [16].

In certain situations or fields (especially in testing), comparisons may be made complicated either by the fact that no reference exists or because no person or organization is taking the initiative of setting up a comparison campaign. In that case, e.g., Shirono *et al.* [17], Thompon *et al.* [11] and Ponomareva *et al.* [19] provide performance evaluation methods for PTs with uncertainty information when there is no reference laboratory available.

In the absence of knowledge of a "true value", the fundamental aim of comparison measurements is to increase as much as possible the confidence level associated with the evaluation of the measurement result. This may be expressed using different terms depending on the branch of statistics involved, but there are basically two common ultimate aims:

- – to obtain a measure of the degree of confidence (or degree of believe) for the differences found in the measured numerical values
- – to obtain a measure of the degree of reliability of the uncertainty evaluations associated to those values.

Note that the attribute "best" often associated to the evaluation of a measurement can only have the meaning associated with (or derived from) the hypotheses under which the analysis is performed, which vary according to subjective preferences.

Without entering into the vast subject of subjectivity vs. objectivity (see, e.g., [19]), it is enough to note here that various degrees of subjectivity are possible, depending on the overall 'level of knowledge'—here a term not necessarily used in the Bayesian sense—associated with a specific measurand.

Strangely, however, this fact is barely referred to, if not totally ignored, in the main approaches to statistics of interest to metrology and testing, i.e., the "error approach" and the more recent "uncertainty approach". Below a roadmap follows to illustrate the role of inter-comparisons in the overall process of gaining progressive knowledge, and the main features of these exercises.

The article maps (in Section 2) the different steps that contribute to the development of knowledge. Looking at a single laboratory, this section explores the existing situation with the choice of method, implementation of a degree of equivalence (DoE), and the necessity to obtain several series of measurements to allow an assessment of repeatability and reproducibility. The interpretation of these concepts, as well as that of accuracy, is the role of the field of metrology and testing. Subsequently, by looking at several laboratories together, we demonstrate that comparison of knowledge is a way of increasing the total level of available knowledge. Section 2 also briefly goes through what a comparison of methods consists of and the advantages this offers. Section 3 presents the principal methods of key inter-comparisons (MRA KC) [1], ISO 5725 [2], and ISO 13528 [3]), and their specificities.

2. HOW KNOWLEDGE IS GAINED IN STEPS

A feature of the measurement process, namely in metrology where it is implicit without always being explicit, is the fact that the acquisition of knowledge always develops with time: knowledge is gained in steps.

2.1. Within-(infra-)Laboratory knowledge

(i) In a single laboratory, the serious job of considering a new measurand is undertaken. First the published/'gray' literature is examined in a search of whatever information may be available on that measurand.

(i-1) If nothing useful is available, proceed to step (ii).

(i-2) If something useful is available, proceed to step (iv): this will be affected by Type B ("other than statistics" method) uncertainty—see also Section 3.2;

(ii) A draft of the "experiment design" is prepared. Almost invariably, the experimental setup (mathematical model ⇒ experimental model, instrumentation) includes in part information already available;

(iii) Case (i-1) does not occur. Otherwise, proceed to step (v);

(iv) Prior to the start of measurement, the measurand may be unknown, and the uncertainty level is affected by partial ignorance regarding the response of the equipment;

(v) If (i-2) applies, the difference from (iii) is that partial ignorance also applies to the measurand-value and uncertainty is too high;

(vi) 1st step of the measurement procedure – A first run generates a series of replicated data. Within this knowledge level (KL) the data can be considered repeated only in the absence of adverse evidence from inference within the knowledge level.

(v-1) If nothing changes in the procedure go to step (vi).

(v-2) Model adjustment can only be possible relative to discrepant behaviour of the equipment: proceed to step (ix);

(vii) 2nd step of the measurement procedure – More runs are replicated in the same laboratory by the same staff—obtaining new series of replicated data—they may be repeated or not, see step (viii) and (ix). If (v-1) applies, the set of two series of data should be considered as repeated, but only after the same scrutiny is performed in (v).

If (v-2) applies proceed to (ix);

However, the possibility of occurrence of between-series significant differences in the representative parameters of the sample distributions (or of the inferred probability distribution) is more likely than for within-series non-random differences between data. More tools are available for this scrutiny, concerning not only the position parameter but also the dispersion parameter: the scrutiny should be considered as the first level of data inter-comparison.

This between-series within-laboratory analysis can be considered as the first level of data comparison, characterized as follows [20].

(viii) *Repeatability*. The set can be formed by several series of data taken over a period of time that is much longer than the "short period" indicated for "repeatability" of data to apply. [9]

In metrology, this is the typical case of a standard constructed with the aim of preserving a stable value with time.

In testing, this is the typical aim of a laboratory issuing test results over time under assumed repeatability conditions. The testing case is different, because the test material changes each time, but applies to checks made using a "reference material": however the latter additional information is external to the within-laboratory knowledge, therefore see step (xi). The repeatability condition is assumed obtained by correctly performing the test according to an approved procedure, and this is basically why the result of a test can be obtained as a single value associated with an acceptance limit (tolerance interval), in contrast with the situation in metrology;

(ix) *Reproducibility*. A reproducibility study consists in preparing a "design of experiment" that can obtain sensitivity coefficients for the different influence quantities.

It consists in varying by known amounts each influence factor separately, and checking the overall effect.

These coefficients can also be computed without experimentation by differentiating the model expressed in closed form (analytically): this method may suffer from model imperfections.

The results do not directly inform about the actual variability of an experimental setup in each specific real condition.

The set of results of a run form a single series of non-repeated data [21].

A variability level of the setup should be obtained by performing a specific experimental condition, called "reproducibility condition" (or "intermediate condition" when focused on only specific effects).

It is assumed that an evaluation of reproducibility is achieved, but the truth of this assumption is not particularly easy to check;

(x) *Accuracy*. Measurements are normally assumed to provide a measure of the true value of the measurand, but this assumption cannot be verified. Consistency of results is indifferent to truth; one can be entirely consistent and still be entirely wrong.

Consistent within-laboratory data can only provide a value that may be used as a "laboratory reference value" for internal use, with an associated dispersion that is lower than that of inconsistent data.

Only when a "reference value" is provided as the target value, one can talk about "precision" or "trueness" [2]. However, it can only be a "conventional" or "accepted" reference value, which shows how unavoidable *inter*-subjectivity is. In this respect a within-laboratory reference is, in itself, less reliable than a between-laboratory reference.

For a specific measurand using a specific method, that is the maximum *Knowledge Level* (*KL)* that can be gained within-laboratory (*KL_w*).

At that level, systematic errors/effects can only be inferred from the dispersion level of uncertainty obtained compared with the target uncertainty, but not proved—except when the target value is an accepted (conventional) reference value. Except also, to a limited extent, when the experience accumulated in the laboratory shows the lack of repeatability between the position parameters of subsequent series of data (called "bias between series"), which is typically an instability of the value of the measurand—called "drift" when the change is essentially monotonic with time.

The main factor limiting the value of these within-laboratory comparisons is the fact that the different series are strongly correlated with each other.

The next step toward increasing the *KL* is to compare similar results on the same (or similar) measurand obtained by other laboratories [22], [23].

2.2. Between-(inter-)Laboratory knowledge

(xi) This is knowledge that is *not* available to any of the participants to the comparison prior to the (first) exercise being finalized. Therefore, it is additional to the *KL_w*.
The comparison is supposed to be performed between non-repeated measurements.
The exception to this is in the testing situation, where all participants apply the same standard procedure: this is assumed to generate repeated measurements within a stated uncertainty, an assumption that is valid until there is evidence to the contrary. An inter-comparison can provide such evidence (of a non-standard condition);

(xii) The maximum increase in knowledge from the results of inter-comparison is achieved when all results are fully uncorrelated. The reason for this is that correlated results tend to be less dispersed simply because at least some of the same systematic errors/effects are paired together. Therefore, the degree of correlation must be carefully scrutinized in order to attribute the correct significance to the results;

(xiii) A comparison concerns one of basically two classes of measurands: (a) artefacts, or (b) physical/chemical/ biological/ etc. states. The comparison design should be specific to the class of measurand;

(xiv) The aim of a comparison must be clearly identified, since this affects several aspects of the exercise.

(xv) A chain of comparisons can occur, involving the same participants, either fully or in part;

(xvi) The level of additional knowledge brought in by the participants depends not only on the number of participants but also on the target uncertainty of the exercise or the dispersion of uncertainty levels among the participants;

(xvii) The knowledge level gained in between-laboratory comparisons (*KL_b*) is additional to the *KL_w*, and is the maximum up-to-date level that can be obtained experimentally. Replication of comparisons may compound knowledge, enriching the *KL_b*. However, the wider the range of uncertainties among the participants in the comparison, the less information the comparison outcome can provide about the participants with the lowest uncertainty;

(xviii) From the *KL_b* the participants can gain information about hidden errors in their own realizations: when discrepant results occur, this can help the relevant laboratory to identify systematic errors/effects and correct its model or detect equipment behaviour anomalies.

The *KL_b* is the maximum *KL* that can be gained for a specific measurand from between-laboratory (*KL_w*) exercises. These usually aim to detect differences when using a specific method.

However, there exists the so-called "method-bias", which is generally not detected by inter-comparisons – except, for example, those designed for the specific task of providing a value for a reference material by using different methods.

In these cases, using different approaches/methods can still enhance the level of knowledge. This is particularly important for measurands of fundamental importance. This applies, for instance, to the constants of nature, and allows a critical analysis of the experimental numerical values such as that provided by the CODATA Task Group [24].

2.3. Contribution of diversity: comparison of methods (i.e., of models)

(xix) In order to detect "method bias", [2] the same measurand needs to be subjected to measurements based on different methods.
The comparison between the results is not generally the task of a specific exercise, but happens when a critical review is made of the literature on the specific subject.
A review may typically concern the diversity of experimental approaches or the diversity of approaches to data-analysis of the same experimental method—or both;

(xx) Each experimental approach is based on a different model and involves, either partially or totally, different influence quantities—and consequently different measurement units. Another reason for diversity as regards experimental methodology is the possible division of the measurands into two broad categories that have been demonstrated to require different tools for their analysis: artefacts, and "natural states" (physical, chemical, biological, etc.). Such diversities are a source of possible discrepant results, and can help in identifying missed influence quantities in other methods;

(xxi) The results of any specific experimental method require an analysis, at least in part of a statistical nature;

(xxii) However, the existence of Type B (in the sense of the GUM [25]) components of uncertainty implies that the analysis extends also over the statistical frame. When this type of exercise concerns a specific method, it is the frame of the inter-comparisons (see xi-xviii above). However, very seldom does such an exercise include a substantial diversity of data analysis; most commonly, especially in the testing frame, an agreed single method for the data analysis is used. The commonest statistical diversity factor is between the "frequentist" and the "Bayesian" approaches, which typically give non-identical results because of the different inference methods on which they are based [26]. Another common approach is to consider all inter-comparison data as pertaining to a single population or, less commonly, to a mixture of populations. In actual fact, the diversity can be wider, including non-probabilistic methods;

(xxiii) The above contributions to total *KL* cannot be included either in the *KL_w* or in the *KL_b*, but bring about a level *KL_d*, with higher understanding.

KL_d is the summary of up-to-date knowledge, time remaining the only factor allowing a possible increase of *KL-d* level thanks to replication of some of the above steps—which may invalidate some or all the previous findings.

3. DIFFERENT TYPES OF INTER-COMPARISONS FOR DIFFERENT AIMS

3.1. Direct inter-comparisons

There are basically three different types of between-(inter-)laboratory direct comparison that can be found in general prescriptive documents, each with different aims: (incidentally, GUM [25] is not considered here, not because it is only a Guideline, but because it deals only with within-laboratory single-series data treatment and not with comparisons)

1) MRA Key Comparison (KC): Inter-comparison in *metrology* in the frame of the Mutual Recognition Agreement (MRA), to establish the “degree of equivalence” between National Metrology Institutes [1];

2) ISO 5725: Inter-Comparison to establish the “accuracy (trueness and precision)” of a method in the field of *testing* [2];

3) ISO 13528: Proficiency Test between Laboratories in the field of *testing*, “to determine the performance of individual laboratories for specific tests or measurements, and to monitor the continuing performance of laboratories” [3].

In Table 1 the main features of the methods used are summarized for each type [27].

3.1.1. MRA Key Comparisons (metrology field)

The Key Comparisons were introduced in 1999 as a basic requirement in the text of the Mutual Recognition Arrangement (MRA) by the BIPM, as the end of the very complex process leading to the Calibration and Measurement Capabilities (CMC) of the NMIs. The whole process is pictured in Figure 1: the very beginning of the process relates to the contents of Section 2 of this paper— a full illustration can be found in [21].

KC meaning. The full meaning of a KC remains to some extent controversial. On the one hand, in fact, it may be perceived as a scientific exercise or as a proficiency test.

The text of the MRA Glossary reads: “Key comparison: one of the set of comparisons selected by a Consultative Committee to test the principal techniques and methods in the field (note that key comparisons may include comparisons of representations of multiples and sub-multiples of SI base and derived units and comparisons of artefacts)”. In the Preamble: “… key comparisons carried out using specified procedures which lead to a quantitative measure of the degree of equivalence of national measurement standards”.

It would seem that the above indications are not sufficient to avoid current interpretations of the two meanings of a KC. One interpretation is that a KC is a scientific exercise establishing a quantitative measure of the international degree of equivalence (DoE) of a given standard, as currently maintained in each NMI under its normal – though generally “best” – local working operations and methods.

The other interpretation considers a KC equivalent to a “proficiency test” (PT), a tool of normal use in testing. The definition of a proficiency test, as given in ISO 13528 [3], is as

Table 1. Different types of direct inter-comparisons for different aims: main features of the different methods.

MRA KC [1]	ISO 5725 [2]	ISO 13528 [3]
The aim is to estimate the differences between NMI Standard realizations. MRA Key Comparison (KC): Inter-comparison in metrology in the context of the Mutual Recognition Agreement (MRA), to establish the “degree of equivalence” between National Metrology Institutes	The aim is to estimate the accuracy (trueness and precision) of a specific method in the field of testing	The aim of the proficiency test is to estimate the accuracy and precision of a specific method in the field of testing
The number of participants is generally small and a random choice of participants cannot be assumed. Non-hierarchical	The number of participants, of top level, is generally sufficiently high to allow the assumption of a random choice of participants. Non-hierarchical	The number of participants is generally high. When there are a few participants (e.g.,< 30), the IUPAC/ CITAC recommendation should be followed. Hierarchical
Measurements can be performed using different methods	All participants strictly use the method under test	The same method is used by all participants, as the test concerns the ability to use it correctly
NMIs with different levels of uncertainty (up to > 10:1) can take part in each exercise. The term "bias" is not used	Estimation of the laboratory bias is one of the aims. This bias is assumed to be a random effect	The laboratory bias is estimated
Each local sample is implicitly assumed to represent the local population	Estimation of the method bias is not one of the aims	Homogeneity and stability of the samples are checked
A KCRV is normally – but not always – defined for each exercise, based on the results of (all) participants, usually with an associated uncertainty estimation.	The results of the exercise are the general mean of the test and an estimation of the between-laboratory variance (in addition to the within-laboratory variance)	An assigned value is used, which can be achieved in a variety of ways (reference or consensus)
The difference of each NMI result with respect to the KCRV is called “degree of equivalence” (DoE). The pair differences, called “bilateral degrees of equivalence”, do not require a KCRV	The precision of the model is established	One of several scoring methods can be used to evaluate the results
No result can be considered an outlier, nor discarded, but is noted.	Outlying results are detected and can be rejected	Outlying results are detected and can be rejected

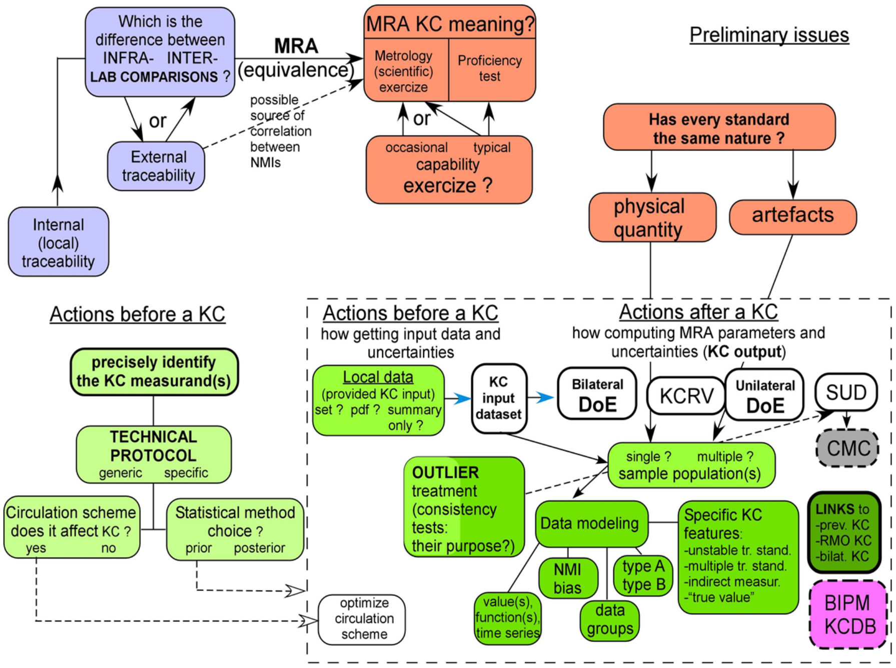
Which is the difference between INFRA- INTER-LAB COMPARISONS ?
MRA (equivalence)
MRA KC meaning?
Metrology (scientific) exercize
Proficiency test
Preliminary issues
or
possible source of correlation between NMIs
External traceability
or
occasional typical capability exercize ?
Has every standard the same nature ?
physical quantity
artefacts
Internal (local) traceability
Actions before a KC
precisely identify the KC measurand(s)
TECHNICAL PROTOCOL generic specific
Circulation scheme does it affect KC ? yes no
Statistical method choice ? prior posterior
optimize circulation scheme
Actions before a KC how getting input data and uncertainties
Actions after a KC how computing MRA parameters and uncertainties (KC output)
Local data (provided KC input) set ? pdf ? summary only ?
KC input dataset
Bilateral DoE
KCRV
Unilateral DoE
SUD
CMC
single ? multiple ? sample population(s)
OUTLIER treatment (consistency tests: their purpose?)
Data modeling
NMI bias
type A type B
value(s), function(s), time series
data groups
Specific KC features: -unstable tr. stand. -multiple tr. stand. -indirect measur. -"true value"
LINKS to -prev. KC -RMO KC -bilat. KC
BIPM KCDB

Figure 1. The process leading to the MRA KC. For acronyms see text, and the following: SUD = systematic unresolved deviations; KCDB = key comparison database; type A, type B: uncertainty components according to GUM (from [21]).

follows: “Proficiency testing by inter-laboratory comparisons is used to determine the performance of individual laboratories for specific tests or measurements, and to monitor the continuing performance of laboratories. The Introduction to ISO 17043 [13] should be consulted for a full exposition of the purposes of proficiency testing. In statistical language, the performance of laboratories may be described by three properties: laboratory bias, stability and repeatability.”

However, proficiency testing should only serve to answer the basic question: can an individual Laboratory continuously demonstrate its ability to correctly conduct a specific procedure? A proficiency test should not be designed to test the *limits* of Laboratory ability to *minimize* its uncertainty, nor to establish how the Laboratory capability ranks within any given community of Laboratories. A proficiency test simply establishes that the Laboratory in question can do its job at its assessed level of competency.

In addition, the term “specified procedures” in the MRA extract quoted above does not refer to a requirement to use a “standard method” in the sense of ISO 5725 [2], the latter associating to it a default uncertainty (and often to a reference value used as the “true value”).

In fact, each NMI participating in a KC is allowed to use different methods and the group of participants can have very different known levels of uncertainty associated to their standards and use different experimental techniques. The KC Protocol is only intended to ensure verification of the transfer standard/s being stable with time (or brought to a specific “ground state” before being passed to the next participant), and verification of the transfer standard/s, or the local standard/s being used in a correct way (consensus best-practice rules).

In addition, the NMIs participating into a KC do not necessarily represent the total population of NMIs, but are only a de facto group. Therefore, the Key Comparison Reference Value (KCRV) prescribed by default by the MRA is not an unbiased estimate of the “best approximation” value of the SI quantity. Consequently, the DoEs relative to the KCRV (unilateral DoEs) are strictly relative to the computed KCRV, regardless of how it is computed. Only the DoEs between pairs of NMIs (bilateral DoE) generally do not suffer from KCRV bias.

On the other hand, the KC represents the only technical basis (and experimental evidence) to allow an NMI adhering to the International Arrangement to accept the degree of equivalence of any other NMI with respect to its own standards. This implies that the measurements leading to that recognition are not occasional (i.e., possibly fortuitous), but represent the limits of the typical capability of a specific NMI.

Therefore, unless a local standard is occasionally in non-“normal” conditions, undetected within the NMI, the measurement data specifically provided by each NMI to a KC (its input data, or in other words, each local sample), is assumed to measure the exact state-of-the-art local standard, i.e. these are its typical values.

The fact that often a limited number of measurements are specifically provided for the KC does not affect the statistical significance of the NMI data provided to the KC. In fact, the limited number of measurements is performed only to check that the standard is at its “normal” operational capability.

The supplied local value is a representative value of the local population; it is not a summary statistic of the (few) specific measurements performed for the KC, any more than is the expected value of the occasional probability density function (PDF) inferred from these specific measurements. It is assumed, instead, to be consistent with the expected value of the local standard as currently maintained, i.e., of the local population of samples normally drawn from that standard. In a word, it is not a “special” value nor is specific to the KC.

Similarly for the associated uncertainty, which is the local capability of realizing the standard. It is not the uncertainty associated to the (few) specific measurements performed for the KC, nor is the second moment of the occasional PDF inferred from these specific measurements. It is evaluated from the local standard as currently maintained, i.e., from the local population of samples normally drawn from that standard. Additionally, only an uncertainty component arising from artefacts arising from the comparison itself (e.g., transfer standard uncertainty, comparison apparatus uncertainty) should be added.

In all instances the KC is a non-hierarchical exercise.

Nature of the standards. Standards can have very different characteristics, depending on the physical or chemical quantity, and they can behave and be used in different ways. These differences reflect on the type of KC [28]. However, as to their intrinsic nature, they can be grouped essentially into two types: (1) artefacts and (2) realization of a physical state or law. This distinction, which should normally generate a substantial difference in the statistical treatment of the data, was introduced back in 2002 and again more recently. Every other characteristic pertains to the specific measurand used in a KC (e.g., single travelling standard) or to an empirical behaviour of a standard realization (e.g., stability with time of the value).

To put it briefly (for more, see [29] and references therein):

(1) Artefact standards (e.g., a piece of metal as a mass standard or as a length standard): a measurement device a "natural" value of which does not exist (device value). Each specific standard carries its own value, though standards can be made with very close values to each other. In the KC statistical treatment, each artefact is regarded as a distinct random variable.

(2) Standards realizing a physical state or law (e.g., phase transition of an ideal substance, vapour pressure law, state related to a fundamental constant, chemical composition of a mixture): a measurement device aimed at accurately realizing the physical state or law. The value of the state or law is unknown but unique for all samples and realizations; it is a physically-based value, and therefore, all realizations aim at approximating the same value. In a KC statistical treatment, each standard is regarded as sampling from the same random variable.

The above has a basic influence on the definition of the measurand of the KC.

Data required as the input to the key comparison. Input data can be required in essentially two different ways. Figure 2 summarizes the case of KCs based on artefact(s):

(i) Using local synthetic input data: this the most usual way, required by most Protocols (upper input type in Figure 2a). Within each NMI, data, x_{ij}, are made available (account being taken of the above discussion about sampling the local standard(s)). The following summary statistics are supplied:

- the expected value y_i of Y_i from the x_{ij} pertaining to X_i, the local random variable, not based only on the x_i, KC data acquired for the KC; Y_i is a new random variable of higher hierarchical rank than X_i
- an uncertainty $u_i(Y_i)$, the variance of the summary variable Y.

(ii) Using the local PDF $\mathcal{F}_i$: it is seldom used (lower input type in Figure 2(a). The statistics of the NMI data are fully conveyed through the PDF, which provides:

- an expectation value $\mathrm{E}(\mathcal{F}(X_i))$, the random variable is X_i;
- an uncertainty u_i, the second moment of the local PDF (accurate only if that can be described by just two moments).

In the case of a KC involving instead the comparison of realizations of a physical state or law, there is a single stochastic variable Q, with an associated single PDF $\mathcal{F}$. All $x_{i,n} \in Q$, where $n = 1 \ldots N$ refer to the NMIs. Therefore, the model of Figure 2 does not apply. However, each NMI contributes to $\mathcal{F}$ with a local $\mathcal{F}_i$, not necessarily equal one to another.

The output data of each participant NMI forms the set of the KC input data. It is analysed by the KC pilot following an agreed statistical method according to a specific model – most often not already specified in the KC Protocol. It enables the obtaining of the KCRV (when this is agreed, as in most cases it is) and the DoEs. Figures 3-4 show typical outcomes of a KC.

The computed KCRV and the DoE (unilateral and bilateral) of each KC are available at the BIPM KCDB, with their links to previous KCs of the same type: the possible SUD are not discussed there, because, the KC being a non-hierarchical exercise, no outliers can be defined nor can the DoE be modified (outliers may only be noted in the Final Report).

This peculiarity does not reflect on the computation of the KCRV, which should be computed with a free-chosen statistical method, but on the representative input values of all participants. There is a line of thinking that instead suggests possibly computing the KCRV on the "maximum consistent set" of the participant results [4] but the effect of such a selection would bias the KCRV, and thus the unilateral DoEs, while the bilateral DoEs are not affected by such a selection.

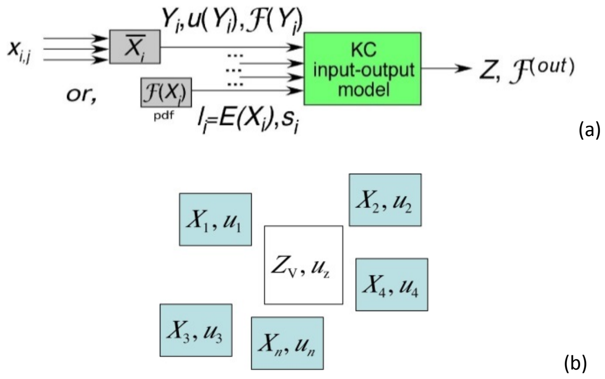

Figure 2. (a) Input-output in a key comparison of artefacts. (b) The output in (a) can be considered as a "virtual artefact", corresponding to an associated new random variable, *Z*. [27].

3.1.2. ISO 5725 Inter-Comparisons (testing field)

The aim of inter-comparisons is indicated in ISO 5725 [2]: "ISO 5725 presents a method of analysis of inter-laboratory comparisons that can be adapted to intra-laboratory comparisons. Users' criteria, concerning operator, equipment, etc. for the tests can be different. In particular, intra-laboratory comparison is realized under specific laboratory conditions. The objective of the ISO 5725 series is: a) to provide useful definitions, b) to provide procedures to assess accuracy (trueness and precision) of measurement methods and results; c) to give guidance and examples to use in practice of trueness and precision data" (emphasis added).

The inter-laboratory comparison aims at validating a single specific method, by means of a non-hierarchical structure and requires top-level laboratories to perform it in order to assess accuracy (trueness and precision) of the method, which can then be used in proficiency tests. In itself "this Standard is not

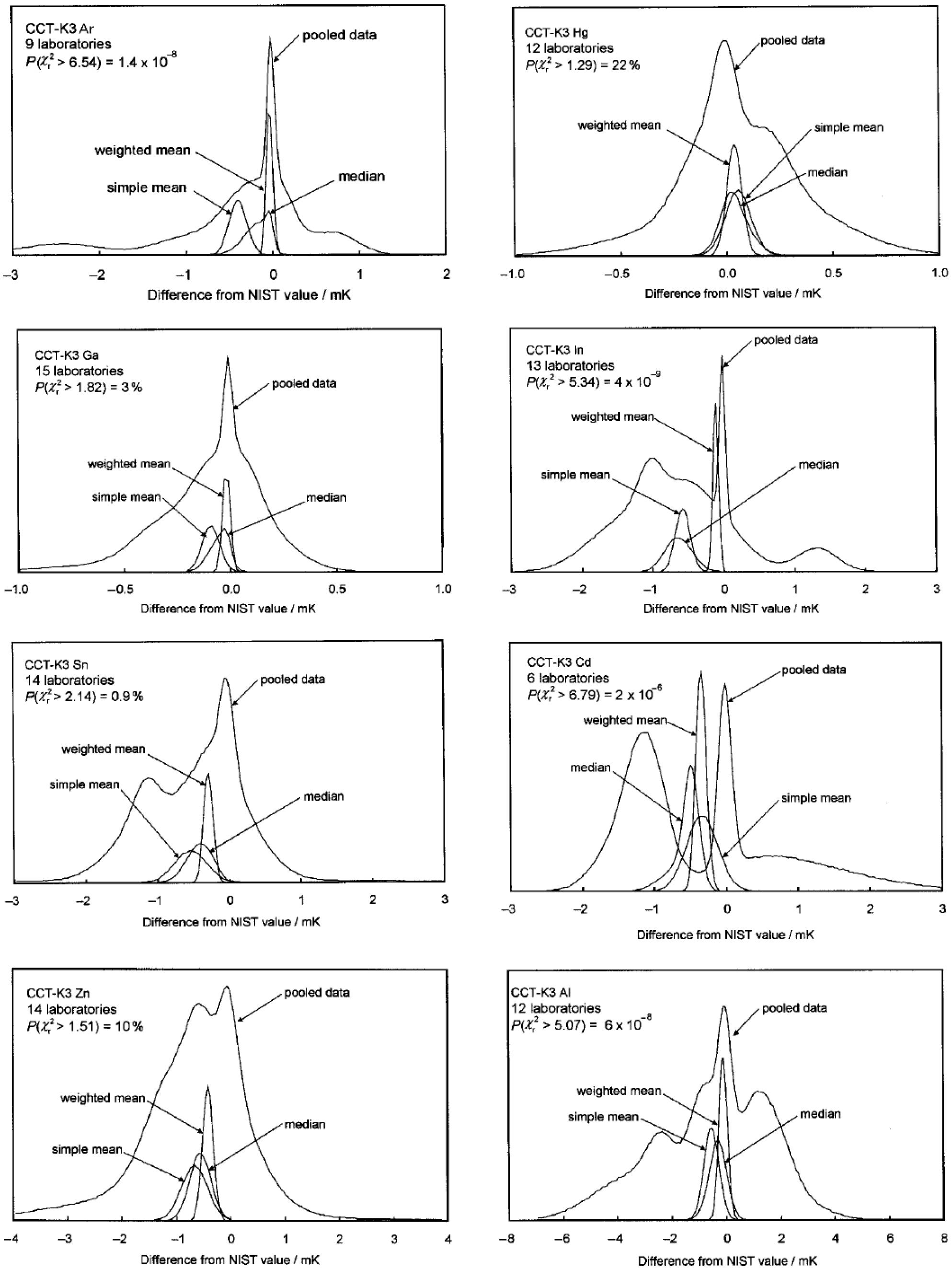

Figure 3. Compound PDF for 8 temperature fixed points of the ITS-90 in a CCT.KCs, obtained as mixtures of several pooled local PDFs [31].

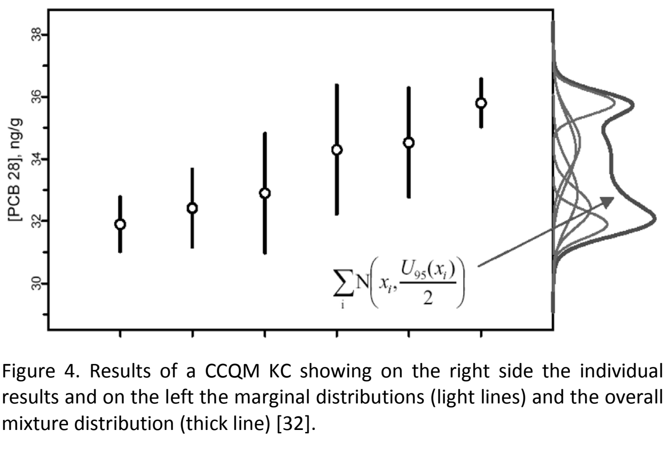

Figure 4. Results of a CCQM KC showing on the right side the individual results and on the left the marginal distributions (light lines) and the overall mixture distribution (thick line) [32].

applicable to proficiency testing or Reference Material Certification". Its use is also restricted, as "the principles of ISO 5725 should also apply to discrete quantities where the quantity is measured on a scale and be considered as continuous variables" (emphasis added).

Accordingly, the number of participants is not so high as in the case of proficiency testing, but it is assumed (without statistical check) that the choice of the participants among the population of the laboratories makes it possible to consider their results sparse at random, so that the Normal distribution is used.

The basic feature of ISO 5725 is that the Normal distribution object of the treatment is that of the expectation provided by the participants, as illustrated in Figure 5. As is clarified in the sentence quoted from ISO 5725, an inter-

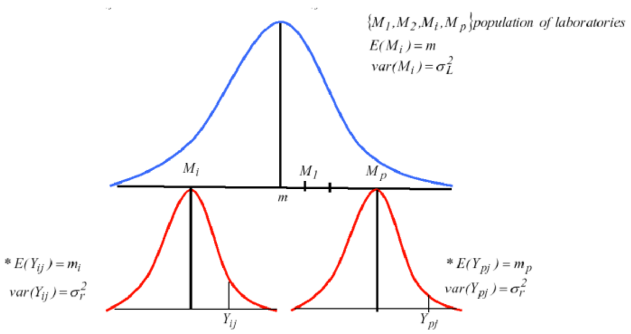

Figure 5. In the lower part the results of each participant are shown: their outcome is the expectation, *E*, of the respective PDF, m_i, and its variance. The upper part represents the distribution of the expectations, M_i, distribution assumed to be Normal: its expectation is *m* and the variance is σ_L^2.

laboratory comparison does not much differ in principle from an intra-laboratory exercise when organized for directly comparing in one place one (or more) standards provided by the different laboratories. In this case, one of the purposes is to establish the differences of the standards of different laboratories.

The ISO 5725 data model is written, where the subscript *i* refer to the *i*-th standard:

$$x_{ij} = a + m_i + \epsilon_{ij} \tag{1}$$

with $i = 1,\ldots, I$; $j = 1,\ldots, J$. See below for the meaning of *a* and m_i.

The term m_i not depending on *j* (i.e., having the same value μ_i for all $j = 1,\ldots, J$) expresses the variability of the *i*-th standard and indicates that part of the model (4) does not apply to all the measurements, but only to the subset concerning the specific *i*-th standard. The index *i* in Equation (1) refers to the *i*-th laboratory, participating with one standard in the exercise ($i = 1,\ldots, I$).

In the calibration frame these exercises admit the use of different methods in different Laboratories, while in the testing frame it is generally not admitted: in fact the "reproducibility" definition mainly differs in this respect between VIM [9] and ISO 5725 [3].

Considering Equation (1), the term *a* in the model has the same meaning that it would have in an intra-laboratory comparison of standards. In the testing frame, *a* is called the general mean (expectation value) in ISO 5725 Standard. When a true value μ can be defined (e.g., an accepted reference value in the sense of ISO 3534 [30], i.e., a known value), in general, $a = \mu + \delta$ where δ is the bias of the measurement method used with respect to the accepted measurement value (ISO 5725).

The term μ_i might have the meaning, $\sum \mu_{in}$, i.e., of the overall effect of the influence parameters in the *i*-th Laboratory. In the testing frame, μ_i is called the "laboratory component bias under repeatability conditions". When a δ exists in this frame, representing the between-laboratory variation, the laboratory bias becomes $\Delta_i = \delta + \mu_i$.

The term ϵ_{ij} in Equation (1) is the random error occurring in every j measurement of the *i*-th participant. ϵ_{ij} is called the "random error occurring in every j measurement of the *i*-th participant under repeatability conditions".

3.1.3. ISO 13528 Proficiency Tests (testing field)

Figure 6 shows a typical outcome of a proficiency test.

ISO 13528 [3] provides detailed descriptions of statistical methods for proficiency testing providers to use to design proficiency testing schemes and to analyse the data obtained from those schemes. It provides recommendations on the interpretation of proficiency testing data by participants in such schemes and by accreditation bodies.

The procedures in ISO 13528 [3] can be applied to demonstrate that the measurement results obtained by laboratories, inspection bodies, and individuals meet specified criteria for acceptable performance [32]. ISO 13528 [3] is applicable to proficiency testing where the results reported are either quantitative measurements or qualitative observations on test items."

The different methods for determining this "standard deviation" for the evaluation of proficiency are contained in the ISO 13528 Standard. Depending on which model is chosen, the standard deviation for the evaluation of proficiency is not an experimental standard deviation in the mathematical sense of the term.

The ISO 13528 Standard suggests five different methods for determining the standard deviation for the evaluation of proficiency as well as eight statistical performance calculations. These choices should be made known to the participants for the interpretation of their performance.

The ISO 13528 Standard can serve as the basis for the implementation of the homogeneity and stability evaluation

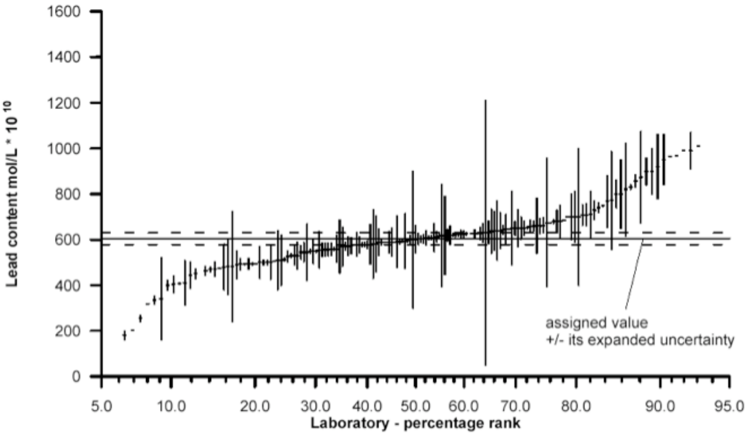

Figure 6. Proficiency test on lead concentration: many more participants laboratories, here shown in the order arising from the obtained numerical result for the same reference material, lead concentration in the matrix [3].

procedures.

The data model proposed in [3] is written as follows:

$$x_i = \mu + \varepsilon_i \qquad (2)$$

where: x_i = Participant result, laboratory *i*, μ = True value for measurand, ε_i = Random error, laboratory *i*.
ISO 5725 [2] and ISO 13528 [3] are complementary and are recommended by ISO 17043 [13].

The proficiency tests are a hierarchical exercise, meaning that the pilot of the PT provides samples from the reference material that is the object of the PT with an assigned target value. The aim of the exercise, differently from that of the one illustrated in the ISO 5725, is to test the current ability of the participants to obtain the reference value.

3.2. Indirect comparisons

Here "indirect" means comparisons lacking a single physical location, for example the result of a search for results in the literature (as in step (i) of the process illustrated in Section 2), where one compares published results without the possibility of checking the comparison outcome by means of a direct comparison.

This case is very common. It is used, for example, by CODATA [24] for the input data of their Least Squares Analysis (LSA) on the numerical values of the physical constants. Only inference about their mutual consistency is possible by means of a critical review and of an expert judgment (Type B component of uncertainty according to the GUM [25]).

In some cases, the measurand is such that, when in doubt, a further direct comparison can be planned and performed (by one of the preceding three methods).

In other cases, this is not possible, for instance for the numerical values of the physical constants. The latter issue is very limiting when it affects the confidence level of the judgment, since it is impossible to adequately check the possible existence of systematic effects that may affect some of the results (e.g., like the current problem with the Planck constant). This issue is postponed to further studies.

4. CONCLUDING REMARKS

The importance and need of inter-comparisons has been illustrated in the frame of the process of progressively increasing the knowledge level, and its main features.

We have also indicated how they can be of different types fulfilling different purposes, by limiting the examples to comparisons being the object of three international codes.

REFERENCES

[1] CIPM MRA-D-05, Measurement comparisons in the context of the CIPM MRA, BIPM Document (https://www.bipm.org/utils/common/documents/CIPM-MRA/CIPM-MRA-D-05.pdf), 2016.

[2] ISO 5725-Part 1 to 6, Accuracy (trueness and precision) of measurement methods and results, 1994.

[3] ISO 13528, Statistical methods for use in proficiency testing by interlaboratory comparison, 2015.

[4] M.G. Cox, The evaluation of key comparison data. An introduction, Metrologia 39 (2002) pp. 589–595.

[5] P. De Courtenay, F. Grégis, The evaluation of measurement uncertainty and its epistemological ramifications, Studies in History and Philosophy of Science 65-66 (2017) pp. 21–32, online June 24 https://doi.org/10.1016/j.shpsa.2017.05.003.

[6] F. Pavese, On the classification in random and systematic effects, AMCTM XI (A.B. Forbes, N.F. Zhang, A.G. Chunovkina, S. Eichstädt, F. Pavese, Eds.), Series on Advances in Mathematics for Applied Sciences vol 89, World Scientific, Singapore, 2018, pp. 58–69.

[7] A.B. Forbes, Approaches to evaluating measurement uncertainty, Int. J. Metrol. Qual. Eng. 3 (2012) pp. 71–77. DOI: https://doi.org/10.1051/ijmqe/2012017.

[8] R. Willink, Confidence intervals and other statistical intervals in metrology, Int. J. Metrol. Qual. Eng., 4 (2013) p. 55–62. DOI: https://doi.org/10.1051/ijmqe/2012029

[9] JCGM 200, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms, 2012.

[10] ISO/IEC 17025, General requirements for the competence of testing and calibration laboratories, 2017.

[11] M. Thompson, S. Ellison, R. Wood, The International Harmonized Protocol for the proficiency testing of analytical chemistries laboratories, Pure Appl. Chem. 78 (2006), pp. 145–196. DOI:10.1351/pac200678010145.

[12] R. Lawn, M. Thompson, R. Walker, Proficiency Testing in Analytical Chemistry, The Royal Society of Chemistry, Cambridge, 1997.

[13] ISO 17043, Conformity assessment – General requirements for proficiency testing, 2010.

[14] ILAC-G13, Guidelines for the requirements for the competence of providers of proficiency testing schemes. 2000. Available online at <http://www.ilac.org/>.

[15] M.G. Cox, The evaluation of key comparison data., Metrologia 39 (2002) pp. 589–596.

[16] C. Elster, B. Toman, Analysis of key comparison data: critical assessment of elements of current practice with suggested improvements, Metrologia 50 (2013) pp. 549–556.

[17] K. Shirono, M. Shiro, H. Tanaka, K. Ehara, Proficiency tests with uncertainty information: Extension of the E_n number for cases with no reference laboratory, Measurement 83 (2016) pp. 135–143.

[18] O. B. Ponomareva, S. Chunovkina, A.V. Shpakov, Testing the proficiency of analytical laboratories by means of inter-laboratory comparison tests – an important element in assurance of the uniformity of measurements, Measurement Techniques (English) 54 (2012).

[19] F. Pavese, On the degree of objectivity of uncertainty evaluation in metrology and testing, Measurement 42 (2009) pp. 1297–1303.

[20] ISO 21748, Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation, 2017.

[21] F. Pavese, A metrologist viewpoint on some statistical issues concerning the comparison of non-repeated measurement data, namely MRA Key Comparisons, Measurement 39 (2006) pp. 821–828.

[22] A.G. Chunovkina, N.D. Zviagin, N.A. Burmistrova, Interlaboratory comparisons, Practical approach for data evaluation, Acta IMEKO 2 (2013) pp. 28–33.

[23] A.G. Chunovkina, A.V. Stepanov, N.A. Burmistrova, Evaluation of inconsistent data: Comparison of two adjustment algorithms, Measurement 91 (2016) pp. 707–712.

[24] http://www.codata.org/about-codata

[25] JCGM, Evaluation of measurement data – Guide to the expression of uncertainty in measurement, BIPM, 2008. https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

[26] R. Willink, Forming a comparison reference value from different distributions of belief, Metrologia 43 (2005) pp. 12–20.

[27] F. Pavese, On hierarchical vs. non-hierarchical comparisons in metrology and testing, Int. J. Metrol. Qual. Eng. 1 (2010) pp. 7–10. DOI: https://doi.org/10.1051/ijmqe/2010004

[28] P. Ciarlini, M.G. Cox, F. Pavese, G. Regoliosi, The use of a mixture of probability distributions in temperature interlaboratory comparisons, Metrologia 41 (2004) pp. 116–121.

[29] F. Pavese, Dependence of the treatment of systematic error in inter-laboratory comparisons on different classes of standards, ACQUAL 1 (2010) pp. 305–315.

[30] ISO 3534, Statistics – Vocabulary and symbols – Part 1 and Part 2, 2006.

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

[32] D.L. Duewer, A Robust Approach for the Determination of CCQM Key Comparison Reference Values and Uncertainties, CCQM-Doc/2004-15, BIPM, Sèvres, 2004.