Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 69 (2014) 263 – 272

Procedia Engineering

www.elsevier.com/locate/procedia

24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013

Verification of Software Applications for Evaluating Interlaboratory Comparison Results

Bojan Acko*, Simon Brezovnik, Boris Sluban

University of Maribor, Faculty of Mechanical Engineering, Smetanova 17, Maribor 2000, Slovenia

Abstract

One of the main tasks of the international metrology system is to assure comparability of measurement results. It can be achieved through recognized traceability chains, which are linking measurement results to references such as measurement units and their realizations. However, the traceability is most commonly limited to measurement instruments and standards of measurement. Since modern measurement applications often use complex software for calculating final results from experimental data, it is very important that all computational links are recognized explicitly and known to be operating correctly. In order to introduce a traceability chain into metrology computation, European project EMRP NEW 06 TraCIM was agreed between EC and European metrology association Euramet. One of the tasks of the project is also to establish random datasets and validation algorithms for verifying software applications for evaluating interlaboratory comparison results. The aims and theoretical approaches of this task are presented in this paper. Background normative documents, calculated statistical parameters, boundary conditions for creating reference data sets, as well as customer interface are described. The verification application will be available on the project web page after finishing the project.

Keywords: traceability; software; interlaboratory comparison; verification; data sets

1. Introduction

Traceability requires that measurement results can be linked to references (such as measurement units) through a documented unbroken chain. If the chain involves computation, as it does in almost all modern measuring systems, it is necessary that all computational links are recognized explicitly and known to be operating correctly [1]. In an

* Corresponding author. Tel.: +386-2-220-7581; fax: +386-2-220-7584.
E-mail address: bojan.acko@um.si

doi:10.1016/j.proeng.2014.02.231

era in which metrology was essentially embodied in hardware, establishing traceability was achieved through a series of calibrations, performed according to documentary standards, using reference artifacts. However, for metrology systems involving significant computation, there are few comparable traceability mechanisms in place. Within the European project EMRP NEW 06 TraCIM we are establishing an infrastructure to allow metadata to be associated with the software component giving a link to the computational aim of the software as specified in a documentary standard, and the date on which the software was last validated using reference data. Then, over the internet, the measuring system will check the latest version of reference data associated with the specified computational aim [1].

One of the tasks of the project is to establish random datasets and validation algorithms for verifying software applications for evaluating interlaboratory comparison results. This task is shared between the Laboratory for Production Measurement at University in Maribor and the German national metrology institute PTB. The decision to include this task in the project was based on extensive research, in which we have examined problems in past international interlaboratory comparisons and availability of tools for evaluating software for statistical calculations. No standards and articles were published in this area so far. The article aims to present general ideas and approaches for establishing an internet–based application, which will serve organizers of different kinds of interlaboratory comparisons to check correctness of their evaluation algorithms based on standardized and other internationally recognized statistical procedures.

2. Interlaboratory comparisons

2.1. Aims and types of interlaboratory comparisons

An interlaboratory comparison is a computationally–intensive metrological tool for evaluating performance of different kinds of metrological laboratories, from national metrology institutes to market-oriented calibration and testing laboratories. Approaches in organization and statistical evaluation of the results can be very different and depend on the aim of an interlaboratory comparison, number of participants, their quality, form of results, etc. Special approaches are used for international key comparisons for evaluating performance quality of national metrology institutes. The application of the procedures to a specific set of key comparison data provides a key comparison reference value (KCRV) and the associated uncertainty, the degree of equivalence of the measurement made by each participating national institute and the degrees of equivalence between measurements made by all pairs of participating institutes [2, 3]. On the other hand, interlaboratory comparisons applied in proficiency testing of testing laboratories follow standardized procedures [4,5] recommending different statistical evaluations of results for different types of interlaboratory comparison and for different ways of reporting measurement results.

2.2. Evaluation of results

In order to evaluate performance of the participants in an interlaboratory comparison, measurement data (with or without associated measurement uncertainties) shall be collected from all the participants and evaluated by means of an agreed statistical approach [2,3,4,5,6]. Single measurement values reported by participants are compared with the agreed assigned (reference) value by considering reported measurement uncertainties and the uncertainty of the assigned value. The basic principle of evaluating performance of participants in an interlaboratory comparison is shown in Fig. 1. Different cases of reporting measurement results shall be considered. In BIPM key or supplementary comparisons and other calibration comparisons, one result and the assigned measurement uncertainty are reported per participant, while in some comparisons in testing participants report more results without uncertainty. The assigned (reference) value can be calculated from the reported measurement values or simply defined as a value of the reference material or as a measurement value of the reference laboratory. The performance metrics depends on the way of reporting results and defining the assigned value. The uncertainty of the assigned value shall be considered in all cases.

Iinterlaboratory comparisons are statistically evaluated by using diverse software, which might produce errors in final results. Error sources could be computational malfunctions, typing mistakes, mistakes in statistical formulae, etc. In order to detect such errors, reference data sets and algorithms for all possible statistical approaches should be

produced and made available to the pilots of interlaboratory comparisons, who are responsible to perform reliable performance metrics. Such reference data sets and calculations shall be cross–checked by using different software packages and by comparisons in different institutes.

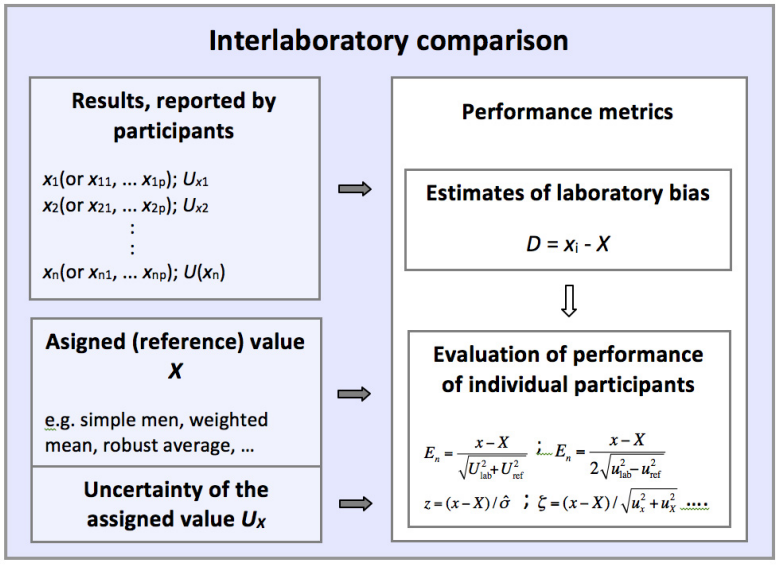

Fig. 1. Evaluation of interlaboratory comparison results.

3. Software tasks to be verified

Input values into the evaluating software are measurement results reported by participants, and corresponding measurement uncertainties, as well as boundary conditions and evaluation strategy. Participants can report one or more results per measurement quantity:

- $x_1, x_2, \ldots x_n$ or
- $x_{11}, x_{12}, \ldots x_{1p}$
  $x_{21}, x_{22}, \ldots x_{2p}$
  ⋮
  ⋮
- $x_{n1}, x_{n2}, \ldots x_{np}$.

The results can be reported with or without measurement uncertainties. The uncertainties can be reported as standard uncertainties (u_{x1}, u_{x2} … u_{xn}) or expanded uncertainties at a certain level of confidence (U_{x1}, U_{x2}, … U_{xn}). Uncertainties are not reported in some cases of comparisons in testing, especially when more than one result is reported per participant.

Output values are calculated in accordance with the users' needs. User can select the set of output values through an intelligent interface. Generated input values are also considering user's boundary conditions. Most common output values in accordance with international standards and recommendations are presented in the following chapters

3.1. Assigned value

The way of determining an assigned (reference) value should be defined prior to the interlaboratory comparison. The value can be determined in advance as a "certified reference value" X_{CRM} (when the material used in a proficiency test is a certified reference material) [4] or a "reference value" X_{RM} (a value of the prepared reference material derived from a calibration against the certified reference values of the CRMs) [4] or a "consensus value from expert laboratories" [4]. However, the most common way of determining the assigned value in calibration interlaboratory comparisons is to calculate it from the reported results x_i as a simple mean [2,6]:

$$x_{ref} = \bar{x} \tag{1}$$

or as a weighted simple mean [2,6]:

$$x_{ref} = \frac{\sum_{i=1}^{n} u^{-2}(x_i) \cdot x_i}{\sum_{i=1}^{n} u^{-2}(x_i)} \tag{2}$$

where:

x_i – measured results reported by participants
u_{xi} – uncertainties of the measured results reported by participants

When the participating laboratories report more than one measured value without stating uncertainty of measurement (in proficiency testing of testing laboratories), the reference value is calculated as a "robust" average [4]:

$$x_{ref} = \sum x_i^* / p \tag{3}$$

where:

$$x_i^* = \begin{cases} x^* - \delta, & if\ x_i < x^* - \delta \\ x^* + \delta, & if\ x_i > x^* + \delta \\ x_i & otherwise \end{cases}$$

x^* = median of x_i $\quad (i = 1, 2, \ldots, p)$

$\delta = 1{,}5 s^*$

$x_1, x_2, \ldots x_p$ – items of data, sorted into increasing order

3.2. Uncertainty of the assigned value

The assigned value is always determined experimentally with some uncertainty. The uncertainty depends on the way of determining the assigned value and is calculated by following standardized or internationally recognized procedures [2,3,4,6,7]. If the assigned value is calculated as a simple mean, its standard uncertainty is [2,6]:

$$u(x_{ref}) = \sqrt{\frac{\sum u^2(x_i)}{n}} \tag{4}$$

where:

$u(x_i)$ – uncertainties reported by participating laboratories
n – number of participants (reported uncertainties)

Standard uncertainty of the weighted mean is [2,6]:

$$u(x_{ref}) = \frac{1}{\sqrt{\sum_{i=1}^{n} u^{-2}(x_i)}} \tag{5}$$

where:

$u(x_i)$ – uncertainties reported by participating laboratories
n – number of participants (reported uncertainties)

The following standard uncertainty is assigned to the robust average [4]:

$$u(x_{ref}) = 1{,}25 \cdot s^* / \sqrt{p} \;\; ; \;\; u_i \text{ not reported} \tag{6}$$

$$u(x_{ref}) = \frac{1{,}25}{p} \sqrt{\sum_{i=1}^{p} u_i^{\,2}} \;\; ; \;\; u_i \text{ reported} \tag{7}$$

where:

$$x_i^* = \begin{cases} x^* - \delta, & if \;\; x_i < x^* - \delta \\ x^* + \delta, & if \;\; x_i > x^* + \delta \\ x_i & otherwise \end{cases}$$

x^* = median of x_i $(i = 1, 2, \ldots, p)$

$\delta = 1{,}5 s^*$

$x_1, x_2, \ldots x_p$ – items of data, sorted into increasing order

$$s^* = 1.134 \sqrt{\sum (x_i^* - x^*)/(p-1)}$$

If the assigned value is defined by perception, the following standard deviation is assigned to it [4]:

$$\hat{\sigma} = \sqrt{(\phi \times \sigma_L)^2 + (\sigma_r^{\,2}/n)} \tag{8}$$

where:

$\sigma_L = \sqrt{\sigma_R^{\,2} - \sigma_r^{\,2}}$
σ_R – reproducibility standard deviation
σ_r – repeatability standard deviation
n – number of replicate measurements each laboratory is to perform
$\sigma_R = 0{,}02 c^{0{,}8495}$
c – concentration of chemical species to be determined in percent (mass fraction)

In the case of defining the assigned value from the results of a precision experiment, the standard deviation is expressed as [4]:

$$\hat{\sigma} = \sqrt{\sigma_L^{\,2} + (\sigma_r^{\,2}/n)} \tag{9}$$

where:

$\sigma_L = \sqrt{\sigma_R^{\,2} - \sigma_r^{\,2}}$
σ_R – reproducibility standard deviation
σ_r – repeatability standard deviation
n – number of replicate measurements each laboratory is to perform

3.3. Performance statistics

Performance statistics is used for evaluating performance of participating laboratories. The final result for single laboratory is most usually "passed" or "failed". Corrective actions shall be taken by the laboratory, which fails the interlaboratory comparison. The first step in the performance statistics is to evaluate estimates of laboratory bias. An estimate could be evaluated as an absolute difference [4,6]:

$$D = x - X \tag{10}$$

where:

x – result reported by a participant
X – assigned value

or as a percentage difference [4,6]:

$$D_{\%} = 100 \cdot (x - X)/X \tag{11}$$

The laboratory bias is than used in different types of evaluation parameters, which should be within certainty limits in order to pass the comparison.

z–score is used in proficiency testing, where no uncertainties are reported by participating laboratories. The *z*–score is calculated by the following equation [4]:

$$z = (x - X)/\hat{\sigma} \tag{12}$$

where:

x – result reported by a participant
X – assigned value
$\hat{\sigma}$ – standard deviation for proficiency assessment

When a participant reports a result that gives rise to a *z*-score above 3,0 or below −3,0, then the result shall be considered to give an "action signal". Likewise, a *z*-score above 2,0 or below −2,0 shall be considered to give a "warning signal". A single "action signal", or "warning signals" in two successive rounds, shall be taken as evidence that an anomaly has occurred that requires investigation.

E_n numbers are used in the comparisons, in which participating laboratories report measurement uncertainties in accordance with the *Guide to the expression of uncertainty in measurement* (GUM). If the reference value X is calculated as a simple mean, the following equation is used [2,3,4,6]:

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{ref}^2}} \tag{13}$$

where:

U_{ref} – expanded uncertainty of the reference value X
U_{lab} – expanded uncertainty of a participant's result x

If the reference value X is calculated as a weighted mean, the E_n value is calculated as follows [2,3,6]:

$$E_n = \frac{x - X}{2\sqrt{u_{lab}^2 - u_{ref}^2}} \tag{14}$$

where:

u_{ref} – standard uncertainty of the reference value X
u_{lab} – standard uncertainty of a participant's result x

When uncertainties are estimated in a way consistent with the Guide to the expression of uncertainty in measurement (GUM), E_n numbers express the validity of the expanded uncertainty estimate associated with each result. A value of $|E_n| < 1$ provides objective evidence that the estimate of uncertainty is consistent with the definition of expanded uncertainty given in the GUM.

z'–scores are used when the assigned value is not calculated using the results reported by the participants and when the participants don't report uncertainties of their results. The *z'*–score is calculated by the following equation [4]:

$$z' = (x - X) / \sqrt{\hat{\sigma}^2 + u_x^2} \tag{15}$$

where:

x – result reported by a participant
X – assigned value
$\hat{\sigma}$ – standard deviation for proficiency assessment
u_X – standard uncertainty of the assigned value X

z'–scores shall be interpreted in the same way as z–scores using the same critical values of 2,0 and 3,0.

ζ–scores are used when the assigned value is not calculated using the results reported by the participants and when the participants report uncertainties of their results. The ζ –score is calculated by the following equation [4]:

$$\zeta = (x - X) / \sqrt{u_x^2 + u_x^2} \tag{16}$$

where:

u_x – laboratory own estimate of the standard uncertainty of the result x
u_X – standard uncertainty of the assigned value X

When there is an effective system in operation for validating laboratories' own estimates of the standard uncertainties of their results, ζ-scores may be used instead of z-scores, and shall be interpreted in the same way as z-scores, using the same critical values of 2,0 and 3,0.

Another criteria are E_z-score [4]. Both values E_{z-} and E_{z+} shall be between -1 and 1 in order to be able to claim thet the participating laboratory performance is satisfactory.

$$E_{z-} = \frac{x - (X - U_x)}{U_X} \text{ and } E_{z+} = \frac{x - (X + U_x)}{U_X} \tag{17}$$

where:

x – result reported by a participant
X – assigned value
U_X – expanded uncertainty of X
U_x – expanded uncertainty of x

3.4. Additional parameters

Additional parameters to be evaluated by interlaboratory comparison software are different kinds of significance tests, confidence elipse, rank correlation test and repeatability standard deviations [4].

4. On-line software validation application

The interlaboratory comparison software validation application allows the user to define boundary conditions for creating random data sets, for which selected reference statistical quantities are calculated. The user's software is then validated by comparing reference quantities with those calculated by the user's software.

4.1. User's interface

The user's interface consists of three modules:

- Selection of boundary conditions for creating data sets,
- Selection of statistical quantities to be calculated,
- Computation of statistical quantities and graphical presentation.

The first module (Fig. 2) is offering the customer to define the following interlaboratory comparison characteristics:

- Number of participants, which is not limited,
- Information about reporting uncertainty of measurement,
- Number of results reported by single participant,
- Target value for the reported result (normally nominal value of the measurand),
- Variation of results (this value can be extracted from real interlaboratory comparison results),
- Accuracy of results (number of decimal places is selected based on the knowledge about real interlaboratory comparison results),
- Type of measurement uncertainty (standard or expanded; the coverage factor can be selected in the case of expanded uncertainty),
- Variation of the measurement uncertainty (this value can be extracted from real interlaboratory comparison results),
- Accuracy of measurement uncertainty (number of decimal places is selected based on the knowledge about real interlaboratory comparison results).

When all boundary conditions are defined, a random data set is created. This data set contains all numerical characteristics of real interlaboratory comparison results. Generation of random data sets can be repeated unlimited number of times. After the data set is generated, the user can select statistical quantities (section 3) to be verified. In the final module, the customer can see reference results and their graphical presentation.

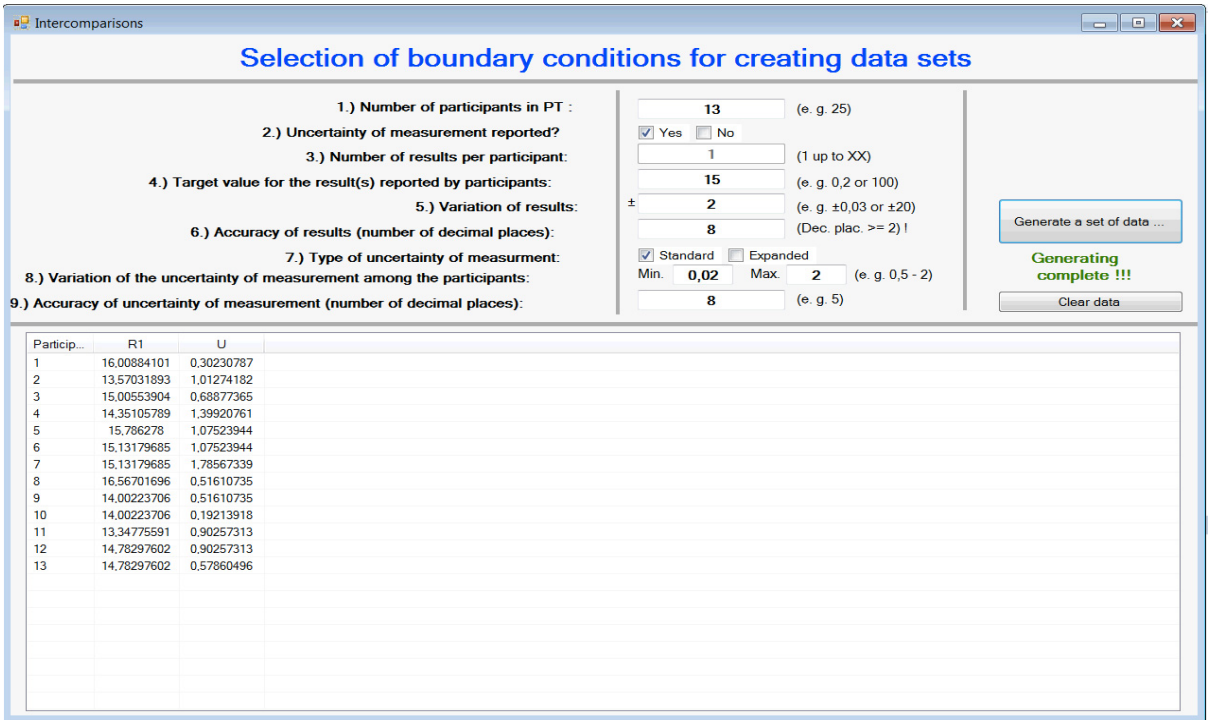

Fig. 2. Module for selecting boundary conditions for creating data sets.

5. Conclusion

The main purpose of interlaboratory comparisons is to give reliable information on participants' competences and metrological capabilities. However, the results can give a completely wrong picture about the participants, if wrong performance metrics is chosen or if unreliable software is used for statistical calculations. Therefore, all algorithms and calculations shall be validated before their use in such delicate evaluation procedure. For this purpose, we are developing an on-line validation tool. The universal on-line application for validating different software packages for interlaboratory comparison data calculation is planned to be a free accessible internet application, which is aimed to serve organizers of all interlaboratory comparisons, which are following standardized or internationally recognized rules. The application is still being developed in the frame of running EMRP TraCIM project. The first module for selecting boundary conditions for creating data sets has already been finished and agreed among the participants in the corresponding project work package. After finishing the second and the third module, the calculation results will be validated by means of using different kinds of software and by comparison among the project participants. The main purpose of using the presented application will be to avoid misinterpretations of interlaboratory comparison results that might lead to wrong evaluation of the participants' performance capability. Therefore, the application will help to improve international comparability and traceability of measurement results.

Acknowledgements

The authors would like to acknowledge funding of the presented research within the European Metrology Research Programme (EMRP) in the Joint Research Project NEW06 TraCIM. Furthermore, fruitful professional discussions within the research group, especially with Phisikalisch Technische Bundesanstalt, are highly appreciated.

References

[1] A. Forbes, NEW06 TraCIM – Traceability of computationally - intensive metrology, EMRP JRP Protocol, NPL, Teddington, 2012.
[2] F. Härtig, K. Kniel, Critical observations on rules for comparing measurement results for key comparisons, Measurement 46 (2013) 3715–3719.
[3] M. G. Cox, The evaluation of key comparison data, Metrologia 39 (2002) 589–95.
[4] ISO 13528, Statistical methods for use in proficiency testing by interlaboratory comparisons, ISO copyright office, Geneva, 2005.
[5] ISO 5725-1: Accuracy (trueness and precision) of measurement methods and results– Intermediate measures of the precision of a standard measurement method, ISO copyright office, Geneva, 2001.
[6] B. Acko, Calibration of line scales – EUROMET Key Comparison L-K7: final report, Metrologia 49 (2012) technical supplement.
[7] S. Raczynski, Uncertainty, Dualism and Inverse Reachable Sets, Int. Journal of Simulation Modelling 10 (2011) 38-45.