Int j simul model 14 (2015) 2, 227-237
ISSN 1726-4529 Original scientific paper

VERIFICATION OF STATISTICAL CALCULATIONS IN INTERLABORATORY COMPARISONS BY SIMULATING INPUT DATASETS

Acko, B.[*]; Brezovnik, S.[**]; Crepinsek Lipus, L.[*] & Klobucar, R.[*]
[*] University of Maribor, Faculty of Mechanical Engineering, Smetanova 17, 2000 Maribor, Slovenia
[**] Gorenje gospodinjski aparati, d. d., Partizanska 12, 3503 Velenje, Slovenia
E-Mail: bojan.acko@um.si, simon.brezovnik@gmail.com, lucija.lipus@um.si, rok.klobucar@um.si

Abstract
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. Statistical analysis of different types of interlaboratory comparisons is based on numerous statistical quantities, which depend on the form of results reported by the participants, way of determining assigned value and its uncertainty, outliers etc. Complex statistical analysis is normally performed by using different software applications. In order to check the performance of those applications, we have developed validation software, which consists of a user interface, data generator and a module for calculating all standardised statistical quantities used for evaluating interlaboratory comparison results. This article is presenting our approach for verifying the validation software before offering it to the metrology community. The verification is based on simulating different cases of interlaboratory comparisons and on comparing results of statistical calculations between different SW packages like Wolfram Mathematica and Microsoft Excel.
(Received in March 2014, accepted in January 2015. This paper was with the authors 5 months for 1 revision.)

Key Words: Interlaboratory Comparison, Validation Software, Performance Metrics, Verification, Simulation

1. INTRODUCTION

Quality of a measurement result is most commonly expressed with an interval of possible scattering around the true value of a measured quantity. In calibration and in advanced measurements, this interval is expressed in terms of measurement uncertainty and is determined according to [1]. However, in most industrial measurements and in some conformity assessment activities like testing and inspection, the measurement result is not accompanied with the information about measurement uncertainty. The interval of allowed scattering is defined with a tolerance zone or expected level of accuracy.

Regardless to the way of expressing measurement result quality, laboratories in metrology business like calibration and testing laboratories are obligated to prove their performance by taking part in interlaboratory comparisons [2-4]. National metrology institutes having their calibration and measurement capabilities (CMCs) published in the international Key Comparison Database (KCDB) must take part in key and supplementary BIPM comparisons on regular basis in order to prove their performance capability on highest metrological level [5-7]. Interlaboratory comparisons are also very important for proving quality of measurements, where measurement uncertainty is hard to determine [8]. In testing, laboratory bias may be assessed by tests on reference materials, when these are available. Otherwise, interlaboratory comparisons provide a generally available means of obtaining information about laboratory bias. However, stability and repeatability will affect data obtained in comparison, so that it is possible for a laboratory to obtain data in a round of a proficiency test which indicate bias that is actually caused by poor stability or poor repeatability [3].

2. SOFTWARE FOR VALIDATING STATISTICAL CALCULATIONS IN INTERLABORATORY COMPARISONS

2.1 Backgrounds

European research project EMRP NEW 06 TraCIM [9] aims to establish random datasets and validation algorithms for verifying statistical calculations in interlaboratory comparisons. This task is shared between the Laboratory for Production Measurement at University in Maribor and the German national metrology institute PTB.

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. [10, 11]. In order to enable validation of statistical calculations in all possible types of interlaboratory comparisons, we have developed a special software application [12, 13] that allows the user to choose between different types of comparisons in calibration and testing. The user can select boundary conditions for generating data sets and statistical quantities to be calculated. The software is offering different selection possibilities by following general rules for comparisons in calibration [2, 3, 10, 11] and in testing [4].

2.2 Statistical quantities

Statistical quantities that are calculated in an evaluation process of intercomparison results are described in detail in [4, 12-15]. Only a short overview is given here in order to establish a statistical background for the software verification process.

Assigned (reference) value

The assigned value can be calculated as a simple mean (eq. (1)), weighted mean (eq. (2)) or a "robust average" (eq. (3)).

$$X = \bar{x} \tag{1}$$

$$X = \frac{\sum_{i=1}^{p} u^{-2}(x_i) \cdot x_i}{\sum_{i=1}^{p} (x_i)} \tag{2}$$

where:

x_i – measured results reported by participants

$u(x_i)$ – uncertainties of the measured results reported by participants

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

where:

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

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

Uncertainty of the assigned (reference) value

Standard uncertainty of the simple mean is calculated by eq. (4), of weighted mean by eq. (5) and of robust average by eqs. (6) or (7). If the assigned value is defined by perception or from a precision experiment, standard uncertainty of the assigned value is not calculated. Only standard deviation is calculated by eq. (8).

$$u(X) = \sqrt{\frac{\sum u^2(x_i)}{p}} \tag{4}$$

where:
p – number of participants

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

$$u(X) = 1.25 \cdot \frac{s^*}{\sqrt{p}} \; ; \; u(x_i) \text{ not reported} \tag{6}$$

$$u(X) = \frac{1.25}{p} \sqrt{\sum_{i=1}^{p} u(x_i)^2} \; ; \; u(x_i) \text{ reported} \tag{7}$$

where:

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

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

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

Performance statistics

The first step in the performance statistics is to evaluate estimates of laboratory bias. An estimate could be evaluated as an absolute difference by eq (9) or as a percentage difference by eq (10).

$$D = x - X \tag{9}$$

$$D_{\%} = 100 \cdot (x - X)/X \tag{10}$$

where:
x – result reported by a participant
X – assigned value

In the second step, performance evaluation criteria are calculated. Selection of proper criterion depends on the case of an interlaboratory comparison and is described in [4, 12-13].

z-score:

$$z = (x - X)/\hat{\sigma} \tag{11}$$

where:
$\hat{\sigma}$ – standard deviation for proficiency assessment

E_n number (eq. (12) is used when x and X are not correlated results, while eq. (13) is used when x and X are correlated)**:**

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{ref}^2}} \tag{12}$$

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

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

z'–score:

$$z' = (x - X)/\sqrt{\hat{\sigma}^2 + u(X)^2} \tag{14}$$

where:
$u(X)$ – standard uncertainty of the assigned value X

ζ–score:

$$\zeta = (x - X)/\sqrt{u(x)^2 + u(X)^2} \tag{15}$$

where:
$u(x)$ – laboratory own estimate of the standard uncertainty of the result x

E_z-score:

$$E_{z-} = \frac{x - (X - U(x))}{U(X)} \text{ and } E_{z+} = \frac{x - (X + U(x))}{U(X)} \tag{16}$$

where:
$U(x)$ – expanded uncertainty of the result x
$U(X)$ – expanded uncertainty of the assigned value X

3. SIMULATED DATASETS FOR TESTING CALCULATION OF STATISTICAL QUANTITIES

Created datasets aim to simulate real cases and some extreme situations of reporting measurement results in interlaboratory comparisons in calibration and testing. Typical datasets contain dimensionless real numbers representing reported measurement results:

- x_i, $i = 1, \ldots, p$

with or without measurement uncertainties. Measurement uncertainties are simulated as dimensionless real numbers representing standard or expanded uncertainties:

- u_i, $i = 1, \ldots, p$
- U_i, $i = 1, \ldots, p$; $U = k \cdot u$.

The expansion coefficient k for calculating expended measurement uncertainty will be in all cases $k = 2$, which corresponds to the confidence probability of approx. 95 %.

The following cases will be simulated:

- Small number of participants (5),
- Big number of participants (50),
- Small accuracy of reported results (1 decimal place),
- High accuracy of reported results (10 decimal places),
- All results very similar or equal, reported uncertainties with small variations,
- All results very similar or equal, reported uncertainties with big variations,
- Big variation of measurement results but no significant outliers,
- Small variation of measurement results with 1 or 2 significant outliers.

Table I: Selected combinations of simulated datasets and tested statistical quantities.

Dataset	Assigned value	Uncertainty of assig. value	Deviation of the scheme	Bias	Evaluation criterion
x_i u_i	Simple mean	yes	no	absolute	E_n
x_i U_i	Simple mean	yes	no	absolute	E_n
x_i u_i	Weighted mean	yes	no	absolute	E_n
x_i U_i	Weighted mean	yes	no	absolute	E_n
x_i u_i	Simple mean	yes	yes	absolute	ζ–score E_z-score
x_i U_i	Weighted mean	yes	yes	absolute	ζ–score E_z-score
x_i	Predetermined (not calcul.)	yes	yes	absolute	z'–score
x_i	Simple mean	yes	yes	absolute	z–score

3.1 Simulated cases of reported intercomparison results

Case 1
Number of reported results: 5
Reported uncertainty: standard
Number of decimal places: 1
Variation of reported results: none
Variation of uncertainties: small

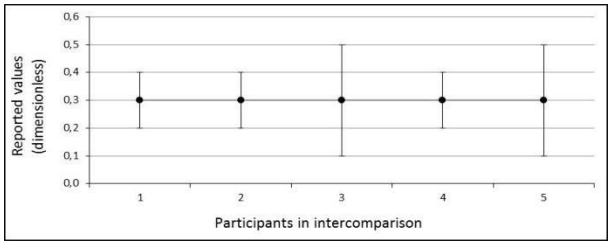

Figure 1: Simulated dataset for case 1.

Case 2
Number of reported results: 5
Reported uncertainty: expanded
Number of decimal places: 3
Variation of reported results: none with 1 outlier
Variation of uncertainties: small

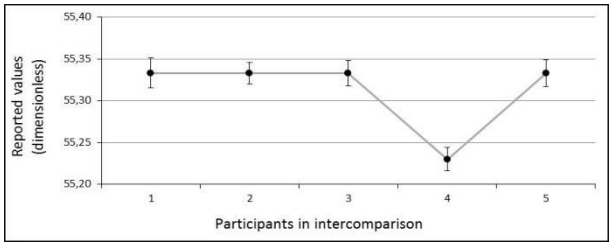

Figure 2: Simulated dataset for case 2.

Case 3
Number of reported results: 10
Reported uncertainty: standard
Number of decimal places: 5
Variation of reported results: big, random
Variation of uncertainties: big, random

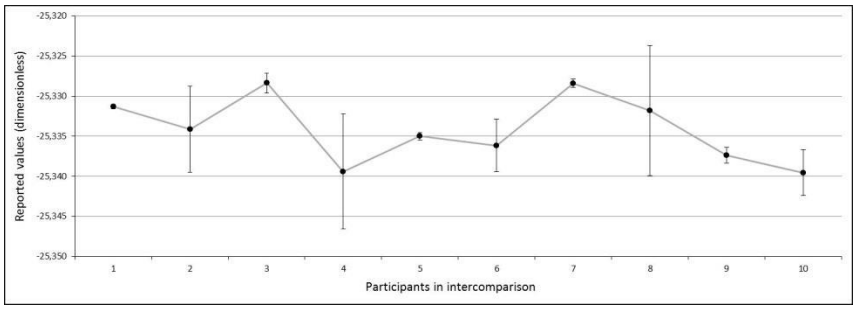

Figure 3: Simulated dataset for case 3.

Case 4
Number of reported results: 20
Reported uncertainty: expanded
Number of decimal places: 5
Variation of reported results: small with 2 outliers
Variation of uncertainties: small with 2 exceptions

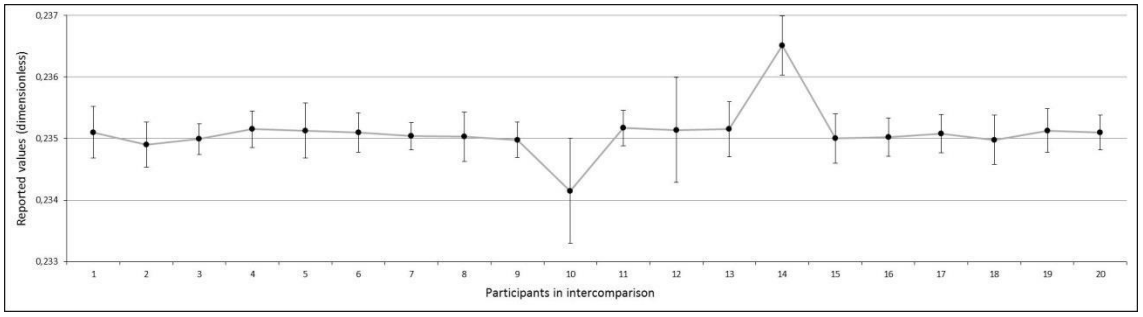

Figure 4: Simulated dataset for case 4.

Case 5
Number of reported results: 20
Reported uncertainty: standard
Number of decimal places: 5
Variation of reported results: small
Variation of uncertainties: big, random

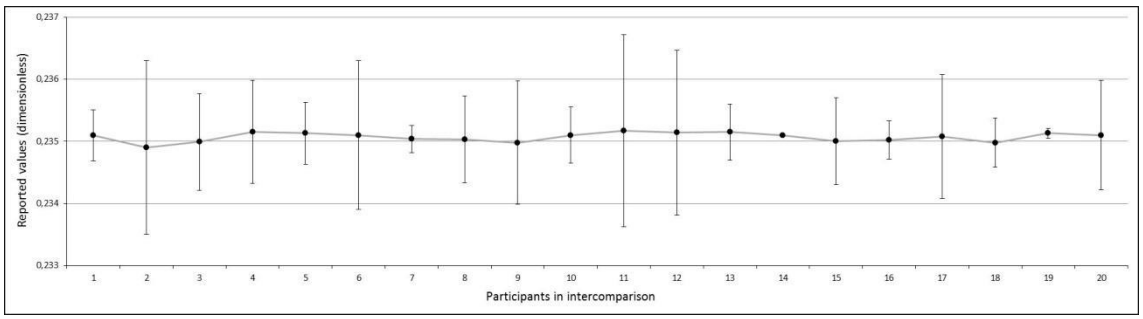

Figure 5: Simulated dataset for case 5.

Case 6
Number of reported results: 20
Reported uncertainty: expanded
Number of decimal places: 3
Variation of reported results: big, random
Variation of uncertainties: big, random

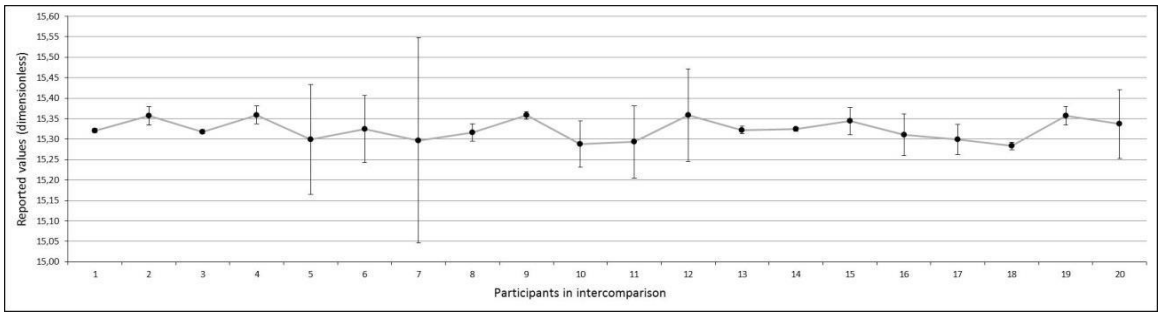

Figure 6: Simulated dataset for case 6.

Case 7
Number of reported results: 50
Reported uncertainty: expanded
Number of decimal places: 10
Variation of reported results: none with 2 outliers
Variation of uncertainties: big, random

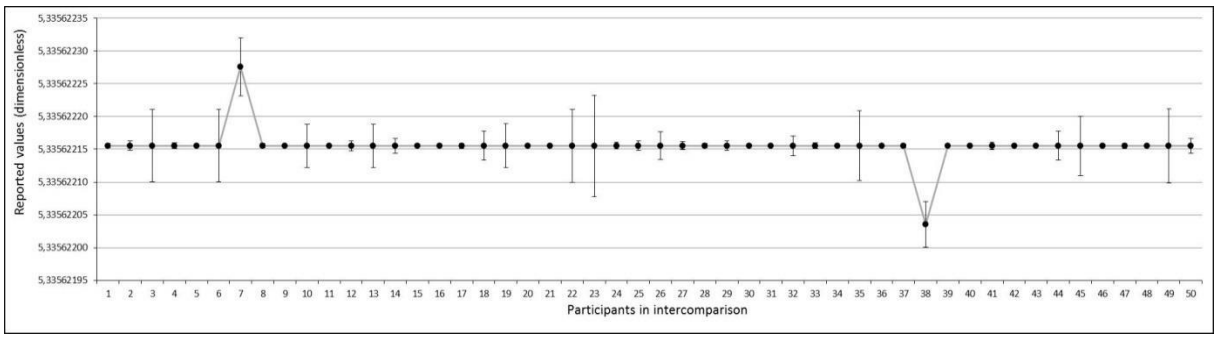

Figure 7: Simulated dataset for case 7.

Case 8
Number of reported results: 50
Reported uncertainty: none
Number of decimal places: 3
Variation of reported results: small
Variation of uncertainties: -

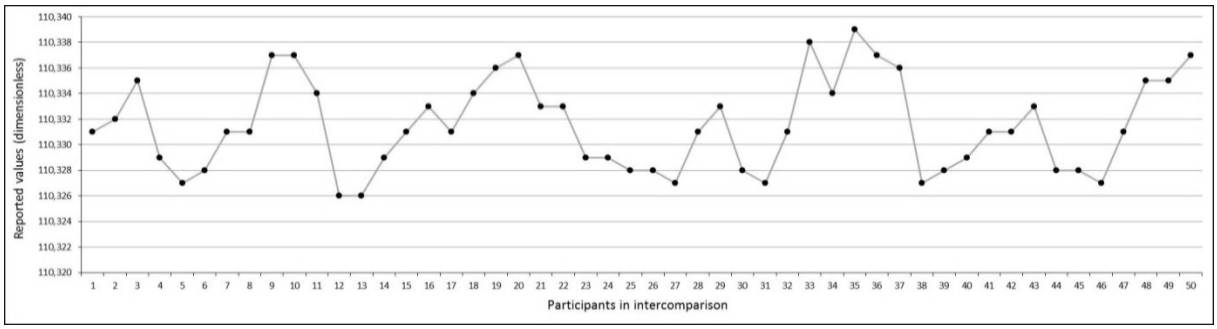

Figure 8: Simulated dataset for case 8.

Case 9
Number of reported results: 10
Reported uncertainty: none
Number of decimal places: 5
Variation of reported results: big, random
Variation of uncertainties: -

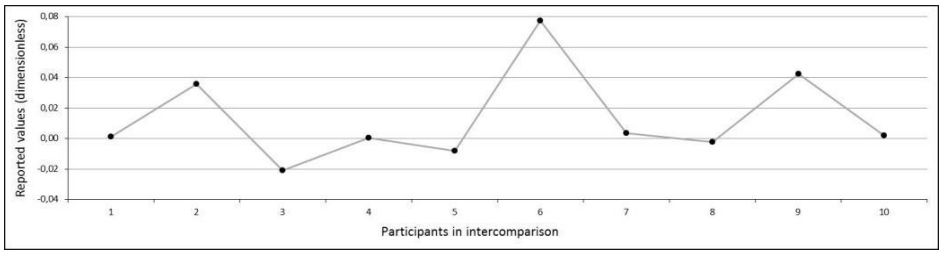

Figure 9: Simulated dataset for case 9.

Case 10
Number of reported results: 10
Reported uncertainty: none
Number of decimal places: 1
Variation of reported results: small
Variation of uncertainties: -

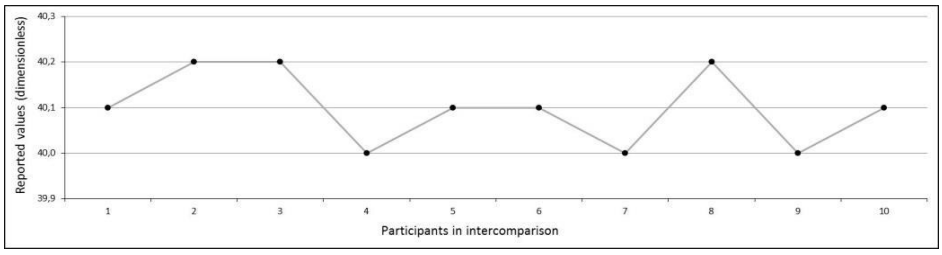

Figure 10: Simulated dataset for case 10.

4. VERIFICATION RESULTS

The verification of statistical calculations in the developed software (created in Visual Studio) was performed by means of comparing results with the results obtained in Wolfram Mathematica and in Microsoft Excel. Ten decimal places were considered as appropriate matching of the results. Ten simulated datasets were used as the input values, but only 3 typical cases will be presented as sample calculations.

4.1 Assigned values and their uncertainties

Verification results for calculated assigned values and their standard uncertainties are presented in Table II. In case 8 laboratories didn't report measurement uncertainties.

4.2 Standard deviations

Standard deviation of the intercomparison scheme (see eq. (8)) was calculated only in cases 8, 9 and 10. In selected case 8, calculated standard deviation was equal in all three SW packages and was $\hat{\sigma} = 0{,}0040091009$.

Table II: Calculated assigned values and measurement uncertainties for datasets 4, 6 and 8.

Data set	SW*	Quantity	Simple mean	Weighted mean	Robust average
4	VS	X	0,2350925000	0,2350892150	-
		$u(X)$	-	0,0000449416	-
	WM	X	0,2350925000	0,2350892150	-
		$u(X)$	-	0,0000449416	-
	ME	X	0,2350925000	0,2350892150	-
		$u(X)$	-	0,0000449416	-
6	VS	X	15,3221874683	15,3221874683	-
		$u(X)$	-	0,0012800318	-
	WM	X	15,3221874683	15,3221874683	-
		$u(X)$	-	0,0012800318	-
	ME	X	15,3221874683	15,3221874683	-
		$u(X)$	-	0,0012800318	-
8	VS	X	-	-	110,3314807639
		$u(X)$	-	-	0,0007087156
	WM	X	-	-	110,3314807639
		$u(X)$	-	-	0,0007087156
	ME	X	-	-	110,3314807639
		$u(X)$	-	-	0,0007087156

*Applied software: VS-Visual Studio, WM-Wolfram Mathematica, ME-Microsoft Excel

4.3 Biases

Only absolute biases (see eq. (9)) were calculated in all three sample cases. Biases were calculated for all participants (see Figs. 4, 6 and 9). In order to make the presentation more clear, only one bias per case is shown. Participant 10 was selected in cases 4 and 6, while participant 25 was selected in case 8. The results are presented in Table III.

Table III: Calculated biases for datasets 4, 6 and 8.

Data set	SW*	Bias
4	VS	-0,0009392150
	WM	-0,0009392150
	ME	-0,0009392150
6	VS	-0,0341874683
	WM	-0,0341874683
	ME	-0,0341874683
8	VS	-0,0034807639
	WM	-0,0034807639
	ME	-0,0034807639

*Applied software: VS-Visual Studio, WM-Wolfram Mathematica, ME-Microsoft Excel

4.4 Evaluation criteria

Evaluation criteria are shown only for one participant per case in this article. Participant 10 was selected in cases 4 and 6, while participant 25 was selected in case 8. The results are presented in Table IV.

Table IV: Calculated evaluation criteria for datasets 4, 6 and 8.

Data set	SW*	E_n number	z - score	ζ - score	E_z - score
4	VS	0,5555944587	-	-2,197664576	-19,9059818192
	WM	0,5555944587	-	-2,197664576	-19,9059818192
	ME	0,5555944587	-	-2,197664576	-19,9059818192
6	VS	0,3055647195	-	-1,219707143	8,5203079877
	WM	0,3055647195	-	-1,219707143	8,5203079877
	ME	0,3055647195	-	-1,219707143	8,5203079877
8	VS	-	-0,8602155854	-	-
	WM	-	-0,8602155854	-	-
	ME	-	-0,8602155854	-	-

*Applied software: VS-Visual Studio, WM-Wolfram Mathematica, ME-Microsoft Excel

4.5 Summary of verification outcomes

Verification of statistical calculations for selected cases of interlaboratory comparisons in calibration and testing was a complex process resulting in a huge number of results. Only a small portion of results is presented in this paper in order to show matching between the selected SW packages. As it can be seen from all presented results, perfect matching is achieved down to 10 decimal places. In fact, the calculations were performed on 15 decimal places. Some discrepancies were obtained in 13th decimal places and lower. However, no real intercomparison cases use more than 10 decimal places and therefore our acceptance criterion was set to this accuracy. It can be confirmed that the criterion was reached in all tested calculations.

5. CONCLUSION

The on-line application for validating calculation software for interlaboratory comparison data calculation is considered to be a free accessible internet application, which is aimed to serve organizers of all interlaboratory comparisons in calibration and testing, which are following standardized or internationally recognized rules. 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. However, such an application should produce very accurate results and shall be able to distinguish between different cases of interlaboratory comparisons. Especially in testing we have very many different situations and the whole statistical calculation process must follow the concrete situation. For these reasons, our SW application went through a complex validation process considering very different cases of reported results with and without corresponding uncertainties. In this testing phase we can confirm correctness of all statistical calculations and selection of evaluation criteria. But this process has not been finished yet. It is planned to put the application in a trial use, during which we intend to collect remarks from users.

ACKNOWLEDGEMENT

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, as well as funding of national research in the frame of the national standard of length by the National Metrology

Institute of Republic of Slovenia (MIRS). Furthermore, fruitful professional discussions within the research group, especially with Phisikalisch Technische Bundesanstalt, are highly appreciated.

REFERENCES

[1] JCGM 100 (2008). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*, 1st edition, Joint Committee for Guides in Metrology
[2] Cox, M. G. (2002). The evaluation of key comparison data, *Metrologia*, Vol. 39, No. 6, 589-595, doi:10.1088/0026-1394/39/6/10
[3] CIPM-MRA (2013). Guide for implementation of the CIPM MRA, CIPM MRA-G-01, v. 1.2, from *http://www.bipm.org/utils/common/CIPM_MRA/CIPM_MRA-G-01.pdf*, accessed on 16-10-2015
[4] ISO 13528 (2005). *Statistical methods for use in proficiency testing by interlaboratory comparisons*, ISO copyright office, Geneva
[5] Mudronja, V.; Barsic, G.; Runje, B. (2014). Calibration of the vertical measuring system of stylus instrument Perthometer S8P, *Technical Gazette*, Vol. 21, No. 1, 141-145
[6] Mudronja, V.; Katic, M.; Simunovic, V. (2014). Realization of the highest level of traceability in Croatian National Laboratory for Length, *Transactions of FAMENA*, Vol. 38, No. 1, 37-44
[7] Lipus, L. C.; Matus, M.; Acko, B. (2013). Optimization of calibrating HeNe laser interferometers by sample-period simulation, *International Journal of Simulation Modelling*, Vol. 12, No. 3, 154-163, doi:10.2507/IJSIMM12(3)2.231
[8] Zbrowski, A.; Matecki, K. (2014). The use of computed tomography to analyse grinding smudges and subsurface defects in roller bearing rings, *Strojniski vestnik – Journal of Mechanical Engineering*, Vol. 60, No. 11, 709-715, doi:10.5545/sv-jme.2014.1817
[9] Forbes, A. (2012). *NEW06 TraCIM – Traceability of computationally-intensive metrology*, *EMRP JRP Protocol*, NPL, Teddington
[10] Cox, M. G. (2007). The evaluation of key comparison data: determining the largest consistent subset, *Metrologia*, Vol. 44, No. 3, 187-200, doi:10.1088/0026-1394/44/3/005
[11] Cox, M. G.; Harris, P. M. (2012). The evaluation of key comparison data using key comparison reference curves, *Metrologia*, Vol. 49, No. 4, 437-445, doi:10.1088/0026-1394/49/4/437
[12] Acko, B.; Brezovnik, S.; Sluban, B. (2014). Verification of software applications for evaluating interlaboratory comparison results, *Procedia Engineering*, Vol. 69, 263-272, doi:10.1016/j.proeng.2014.02.231
[13] Acko, B.; Sluban, B.; Tasic, T.; Brezovnik, S. (2014). Performance metrics for testing statistical calculations in interlaboratory comparisons, *Advances in Production Engineering & Management*, Vol. 9, No. 1, 44-52, doi:10.14743/apem2014.1.175
[14] ISO 5725-1 (2001). *Accuracy (trueness and precision) of measurement methods and results– Intermediate measures of the precision of a standard measurement method*, ISO copyright office, Geneva
[15] Härtig, F.; Kniel, K. (2013). Critical observations on rules for comparing measurement results for key comparisons, *Measurement,* Vol. 46, No. 9, 3715-3719, doi:10.1016/j.measurement.2013.04.079