Accred Qual Assur (2016) 21:91–100
DOI 10.1007/s00769-016-1195-y

CrossMark

GENERAL PAPER

Is the assessment of interlaboratory comparison results for a small number of tests and limited number of participants reliable and rational?

Ewa Szewczak[1] · **Adam Bondarzewski**[1]

Received: 10 November 2015 / Accepted: 22 January 2016 / Published online: 23 February 2016

Abstract Tests and/or test items can sometimes be expensive, unique, or only performed in a few laboratories. There can be cases where assigned values are unknown, there is no information, or only poor information on the probability density function attributed to the test result. Sometimes there are neither reference materials nor the ability to establish consensus values due to a lack of experts. It can be impossible to repeat a test on the same item because it is destroyed during the test itself, or the homogeneity of tested items is unknown and no criteria can be established. Specified technical requirements concerning proficiency testing and interlaboratory comparison schemes are generally not applicable in this situation. However, interlaboratory comparison could allow laboratories to have more confidence in their results. The present paper discusses three statistical methods of assessing interlaboratory comparison results obtained in such conditions. Two methods are based on an assigned value determined from participant results through robust analysis. The third is based on the compatibility of results assessed using the ζ parameter. This paper focuses on an interlaboratory comparison for two laboratories, each testing three samples. The use of statistical methods turns out to be high risk, particularly in terms of falsely accepting results. Additionally, is shown that methods dedicated to small samples are also not efficient in detecting discrepancies of test results.

Keywords Proficiency testing criteria · Quality control of tests · Small samples

Introduction

According to EN ISO IEC 17025 [1] and EA-4/18 [2], accredited laboratories should assure the quality of test results by participating in proficiency testing programs. In the case of a lack of proficiency tests because of, for example, the technical characteristics of the measurement or the low number of existing laboratories in the sector, other methods of assuring quality are accepted. However, interlaboratory comparisons (ILCs) are preferred by accreditation bodies. This is the reason why interlaboratory comparisons are organized often even if there are no reasonable methods of assessing the results.

Typical methods of assessment of ILC results are described in standards EN 17043 [3] and ISO 13528 [4]. Most are based on a known assigned value (*value attributed to a particular quantity and accepted* [4]) and its uncertainty. This knowledge comes from preparing special samples for the purpose of ILC, using certified reference materials (CRMs) or testing the samples at expert laboratories before the ILC. For some statistics used in the assessment of laboratory proficiency, reference laboratories are involved. When it is not possible to apply the above methods, consensus values calculated from participant results using robust analysis are recommended for the estimation of an assigned value. But for a limited number of participating laboratories when *statistical methods become increasingly unreliable*, schemes based on CRMs are preferred in the available literature [5].

However, it is sometimes not possible to apply a recommended method of assessment of ILC results. The

✉ Ewa Szewczak
e.szewczak@itb.pl

[1] Instytut Techniki Budowlanej, Filtrowa 1, 00-611 Warszawa, Poland

assigned value is unknown. Neither are there reference materials nor is there the possibility of establishing consensus values owing to a lack of experts. It is impossible to repeat a test on the same test item because it is destroyed during tests. The homogeneity of tested items is unknown. Moreover, tests and/or test items are expensive or unique, and thus, a small number of tests results are available.

Such situations are frequent in the mechanical testing of construction product conducted to find a characteristic (type) of an unknown product [6]. An example is the mechanical testing of doors, windows, walls, panels, lintels, and small wastewater treatment systems, where both the tests and test items are often expensive. Additionally, in these situations, it is important to assure the quality of the test result because the result can directly affect safety or health. The above problem can also be encountered in laboratories that conduct chemical tests of substance/elements that are rarely presented or expensive and in the medical testing of human tissue.

Performing an ILC test on simplified samples is one of many solutions (e.g., a laboratory that tests the load bearing capacity of small wastewater treatment systems having tanks of about 3 m^3 may take part in an ILC of the compressive strength of concrete blocks of the size order of dm^3), but it does not provide the laboratories and its customers with a sense of security.

Technical requirements specified in EN ISO/IEC 17043 [3] and ISO 13528 [4] or IUPAC Technical Report [5] concerning proficiency testing and ILC schemes are generally not applicable in the situation of interest; i.e., the situation of comparison a small number of laboratories and a small number of samples with no knowledge of the assigned value, when statistical criteria for ILC can only be based on an assigned value and/or standard deviation (SD) taken from the participant. There are commonly used statistical tests of consistency, such as F and t tests, but such statistics seem to be useless in this case because of the high critical values for small samples, which entail a risk of false acceptance. Other statistics (e.g., χ^2) are unsuitable because of the need to know a predetermined value of variance.

The present paper addresses the question: Is it possible to show the reliability of test results and competence of laboratories in an interlaboratory comparison for a small number of possible tests, limited number of participants, no determined assigned value, and no determined permissible uncertainty? Moreover, are statistical assessments of ILC results reliable and rational? This paper considers ILC for two laboratories, each having three samples. This issue has been not considered previously.

Common methods of assessing the consistency of test results

There are three general methods of assessing test results in an ILC:

- assessing the difference between each result and a "true value,"
- comparing laboratory variance (or uncertainty) with predicted, required, or known variance, and
- assessing of comparability of laboratory results.

The last method is the most promising for our purposes because it does not require knowledge of a "true value" or predicted variance.

Typical simple methods of ILC result assessment are described in ISO 13528 [4]. In our case, there is no possibility of establishing reference laboratory, and thus, the E_n number is useless and the z score (z) and *zeta* score (ζ_X in this paper) should be employed instead. These are defined as

$$z = \frac{x - X}{\widehat{\sigma}}, \tag{1}$$

$$\zeta_X = \frac{x - X}{\sqrt{u_{\text{lab}}^2 + u_{\text{av}}^2}}, \tag{2}$$

where x is the participant result, X is the assigned value, $\widehat{\sigma}$ is the SD for proficiency assessment, u_{lab} is the combined standard uncertainty of a participant's result, and u_{av} is the standard uncertainty of the assigned value.

According to Eqs. (1) and (2), both z and ζ_X scores are based on an assigned value (X) and the SD for proficiency assessment ($\widehat{\sigma}$) or standard uncertainty of the assigned value (u_{av}). However, Eq. (2) can be used only if x and X are independent, and therefore, X should not be calculated from the results of participants. Thus, among the statistics listed, only the z score is adopted in this work.

If we assume that the values of X and/or $\widehat{\sigma}$ cannot be determined by any method that is not related to the current comparison, then according to ISO 13528, they should be determined from participant results through robust analysis. It is recommended that Algorithm A [4, 7] be used to obtain robust values of the assigned value and SD. However, the question arises whether this algorithm might be used for the estimation of X and $\widehat{\sigma}$ in the case under consideration, because the intention is not to use the algorithm for a small population of test results.

Robust estimators for small samples were studied by Rousseeuw et al. [8]. Obviously, robustness is not possible for n equal to 1 or 2 (where n is number of results). When $n = 3$ and the location and scale are unknown, it is recommended that the location is estimated as the sample median, but there is no robust scale estimator. For $n \geq 4$,

the authors propose the location be estimated using the *M*-estimator with a smooth ψ function and the median absolute deviation MAD_n using as the auxiliary scale, and analogously, the estimation scale be estimated by the *M*-estimator with a smooth ρ function using med_n (median) as the auxiliary location. In contrast to Algorithm A, functions used for location and scale estimation are monotonic. The question is does the employment of these analyses for the estimation of X and $\widehat{\sigma}$ solve the problem of assessment of ILC for a small number of tests and laboratories?

The estimation of X and $\widehat{\sigma}$ could be avoided using methods of assessment that do not consider an assigned value.

Kacker et al. [9–11] and Kessel et al. [12] considered a discrepancy measure that can be used to check the agreement of test results. They discussed the Birge test, which is a classical test that was developed for checking the consistency of interlaboratory test results, specifically whether measured values might be considered as realizations of a normal probability density function with unknown expected values but known variance [9, 10]. Kacker et al. [11] showed that the Birge test is not consistent with the philosophy of the *Guide to the Expression of Uncertainty in Measurement* (GUM) [13]. The concept of the metrological compatibility of results consistent with VIM3 [14] and GUM has been discussed [11, 12]. According to the VIM3 definition restated in [12], two metrologically comparable results $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$ for the same measurand are said to be metrologically compatible if

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{u(x_1 - x_2)} \leq \kappa, \tag{3}$$

where $[x_i, u(x_i)]$ denotes the measured quantity value and its standard uncertainty, κ is the chosen threshold (conventionally having a value of two). ζ is a function that may be used as a measure of the significance of the difference between two results, $[x_1, u(x_1)]$ and $[x_2, u(x_2)]$. Such a concept of metrological compatibility is consistent with the GUM.

If we assume that measurements of $[x_i, u(x_i)]$ are uncorrelated and their weights are the same, then

$$u^2(x_1 - x_2) = u^2(x_1) + u^2(x_2), \tag{4}$$

and thus,

$$\zeta(x_1 - x_2) = \frac{|x_1 - x_2|}{\sqrt{u^2(x_1) + u^2(x_2)}}. \tag{5}$$

On the above basis, two functions are employed in this paper for the analysis of results of ILC for a small number of laboratories and small number of samples: the ζ function given by Eq. (5) and the z function given by Eq. (1). For the calculation of z, X and $\widehat{\sigma}$ values are determined from participant results through robust analysis. Algorithm A according to ISO 13528 and ISO 5725-5 is used for the calculation of robust $X = X_\mathrm{A}$ and $\widehat{\sigma} = \widehat{\sigma}_\mathrm{A}$, and z_A is calculated as

$$z_\mathrm{A} = \frac{x - X_\mathrm{A}}{\widehat{\sigma}_\mathrm{A}}. \tag{6}$$

Another algorithm, referred to as Algorithm B in this paper and based on robust analysis for small samples following Rousseeuw et al. [8], is employed for the calculation of $X = X_\mathrm{B}$ and $\widehat{\sigma} = \widehat{\sigma}_\mathrm{B}$, and z_B is calculated as

$$z_\mathrm{B} = \frac{x - X_\mathrm{B}}{\widehat{\sigma}_\mathrm{B}} \tag{7}$$

Parameters ζ, z_A, and z_B are then compared in terms of detecting the inconsistency of test results for two laboratories, each testing three samples.

Simulation of interlaboratory comparison

To compare the effectiveness of parameters ζ, z_A, and z_B for small samples, it is considered that two laboratories participate in ILC, and each laboratory performs three tests. During testing, test items are destroyed, and it is thus not possible to repeat a test for the same sample. Three samples of the same product are tested at each laboratory.

This paper takes a single repetition x_{ij} (for $i =$ laboratory 1 or 2 and $j =$ repetition 1, 2, or 3 for each laboratory) as the test result. A relatively wide dispersion of results is assumed. Sources of this dispersion are discussed in the next section.

Simulation of interlaboratory tests is carried out using *Excel Data Analysis Tool*: *Random Number Generation*. The tool is used to generate 12 sets, with each set containing six random numbers drawn from a normal distribution with mean $\mu = 5$ and SD $\sigma = 1$. Such a ratio between the mean and SD is typical for the example of mechanical tests of large items. Each set of six values is then divided into two parts. Each part represents simulated test results (x_{ij}) of one of the two laboratories LAB_i, where $i = 1, 2$.

A discrepancy between results is introduced by introducing $d = 1$ or 2 outliers in the LAB_2 results. The value of an outlier is given by

$$o = x_{2,j} + b, \tag{8}$$

where $x_{2,j}$ is the jth result of laboratory 2 and $b = 2, 3, 4, 5, 10$ is the bias value added to $x_{2,j}$.

In case of three "outliers," which means that all LAB_2 results differ from the results of LAB_1, three random numbers (LAB_2 test results) are drawn from a normal distribution with $\mu_2 = 5 + b$ and $\sigma = 1$. The results of LAB_1 are unchanged. Additionally, to conduct a simulation of two tests performed in two laboratories, the same sets of

data are used but with the exclusion of the third result of each laboratory.

The following three sections present the methods used to assess the simulated results of laboratories.

Method I of assessing the ILC results using the ζ function of the compatibility of test results

Function ζ defined in Eq. (5) requires only knowledge of probability density functions represented by the results of the laboratories $[x_1, u(x_1)]$, $[x_2, u(x_2)]$ and not knowledge of an assigned value. The result for a laboratory conducting n tests (repetitions) is

$$x_i = \frac{\sum_j x_{i,j}}{n}, \tag{9}$$

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

To simplify the problem, we assume that there are the three following main sources of uncertainty $u(x_i)$.

- The characteristic (accuracy) of measuring instruments. Uncertainty is evaluated using data provided by calibration certificates.
- Variability due to repeatability and reproducibility of the test method. Factors affecting this variability depend on the method. In most cases, it is not possible to assess the effect of an individual factor on uncertainty and it is common to use the Type A [13] evaluation of standard uncertainty from the statistical distribution of the values obtained from a series of measurements.
- Variability due to the tested product and its inhomogeneity. The repeatability of the test item is not dependent on the laboratory but on the type of product and its production process.

If it is possible to perform tests on items of known homogeneity or on reference materials, then it is possible to separate variability due to the laboratory from variability due to the tested product. However, in the cases considered here, there is no reasonable way of separating the effects of the tested product and test method on the variability of test results. All historical data concern a small number of tests of different products (tests are expensive, and sample is destroyed during the test). The SD values taken from results obtained in the same laboratory differ appreciably for different types of product, and knowledge of the SD that could be assigned to laboratory uncertainty is thus unavailable. Uncertainty $u(x_i)$ can be estimated only on the basis of the current sample. It seems to be justified, as the only available option in such case, to use the sample SD of current results as an approximation of uncertainty $u(x_i)$ in this article. Hence, in the ζ function (Eq. 5) used as a measure of the difference between the results of two laboratories, we used the mean of the results for laboratory i as x_i and the sample SD of results for laboratory i as $u(x_i)$.

Method II of assessing ILC results using the z score and a robust estimator of the assigned value obtained in Algorithm A according to ISO 13528

To use the z function (Eq. 1), information on the assigned value X and its standard uncertainty is needed. Because there is no reference value and there are no expert laboratories, the calculation of the assigned value has to be based on robust estimation from participant results.

According to Algorithm A, recommended by ISO 13528, the first evaluation of the location X^* and scale s^* estimator is:

$$X^* = \text{med}(x_i) \tag{10}$$

$$s^* = 1.483\ \text{med}|x_i - X^*| \tag{11}$$

where $i = 1, 2, \ldots p$, with p being the number of test results.

Next, estimators are derived through an iterative calculation of X^* and s^*:

$$X^* = \sum \frac{x_i^*}{p},$$

where

$$x_i^* = \left\{ \begin{array}{lll} X^* - 1.5s^* & \text{if} & x_i < X^* - 1.5s^* \\ X^* + 1.5s^* & \text{if} & x_i > X^* + 1.5s^* \\ x_i & & \text{otherwise} \end{array} \right\}, \tag{12}$$

$$s^* = 1.134\sqrt{\frac{\sum (x_i^* - X^*)^2}{p-1}}. \tag{13}$$

An iterative calculation according to ISO 13528 is performed until there is no change from one iteration to the next in the third significant figure of s^* and the equivalent in X^*. Equation (6) is then used for the calculation of z_A, where $X_A = X^*$ and $\widehat{\sigma}_A = s^*$.

The ISO 13528 standard takes the average of all participant measurements of the test material as "result" x_i. In our case, we have only two results x_1 and x_2, referring to LAB_1 and LAB_2, for the calculation of X^* and s^*. Using Algorithm A for $p = 2$ items of data, we always obtain the same z_A (ca. 0.62), regardless of the values of x_1 and x_2, which is of course useless for the assessment of laboratory performance. For this reason, in our calculation results for all tests performed by the two laboratories, x_{ij} ($i = 1, 2$, $j = 1, 2, 3$) replaces x_i in Eqs. (10)–(13) used to estimate X^* and s^* (we then have $p = 6$ values of test results).

Method III of assessing ILC results using the z score and a robust estimator of the assigned value obtained in Algorithm B

According to Rousseeuw et al. [8] for the estimation of X_B and $\widehat{\sigma}_B$, we use the M-estimator of location T_n that is described by

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{x_i - T_n}{S_n}\right) = 0 \tag{14}$$

$$\psi(x) = \frac{e^x - 1}{e^x + 1} = \tanh\left(\frac{x}{2}\right), \tag{15}$$

where T_n is the location estimator and S_n is the scale estimator.

By analogy with Method II, all tests results obtained by the two laboratories x_{ij} ($i = 1, 2, j = 1, 2, 3$) are used as x_i in Eq. (14). The first evaluation X_1 of the location estimator is

$$X_1 = \text{med}\ (x_{i,j}). \tag{16}$$

Next T_n is iteratively calculated using Eq. (14). As recommended by Rousseeuw, T_n is computed using a Newton–Raphson algorithm, the code of which was developed by the author of this work (shown in "Appendix 1").

The scale estimator (median absolute deviation) is calculated as

$$S_n = c_n \cdot 1.483 \cdot \text{med}|x_i - \text{med}(x_i)|, \tag{17}$$

where c_n is a small sample correction factor, dependent on n, which ensures that the median absolute deviation is unbiased [15].

z_B is then calculated according to Eq. (7), where $x = x_i$ is the participant result according to Eq. (9), $X_B = T_n$, and $\widehat{\sigma}_B = S_n$.

Results and discussion

Appropriate interpretation of ζ, z_A, and z_B is necessary to confirm agreement or to alert laboratories of discrepancy after ILC. It is assumed [3, 4] that z scores above 2.0 (or below −2.0) indicate discrepancy. The same critical value is commonly used for ζ [12]. If ζ has a value above 2.0, the difference between test results is deemed significant in view of their standard uncertainties.

Critical values for ζ and z scores should in practice depend on the type of test, tested product, the aim of the test, and other risk factors. They could be derived, for example, from z-based and, t-based uncertainty estimators or an unbiased uncertainty estimator (z/c_4), as has been recommended by Huang even for small samples [16]. However, choice of threshold is not the subject of this article. The main question is are the parameters ζ, z_A, and z_B effective enough in detecting discrepancy between laboratories.

Figure 1 shows the values of ζ, z_A, and z_B obtained for biases $b = 0,\ldots10$ added to the results of LAB_2 (according to the described method of simulation). For one outlier introduced in LAB_2, only a few values of ζ are greater than 1 and none is greater than 2, even for bias of 10 (i.e., 10 multiples of the SD σ). In other words, in this case, ζ has no effectiveness in detecting discrepancies. For each bias $b = 0$, 2, and 3, one ζ value exceeds 1, but for $b = 0$ this should be interpreted as a false signal.

Better results concerning detection of discrepancies are obtained for z_A and z_B parameters.

A similar situation occurs for two outliers introduced in LAB_2 results, but ζ becomes more effective and z_A and z_B a little less effective.

There is a notable change in the case of three outliers (Fig. 1c; Table 1). In this case, only ζ detect discrepancy of the tests results, while z_A and z_B do not. In fact, three outliers in LAB_2 correspond to the situation that all the results of LAB_2 are incompatible with the results of LAB_1 and this means that the laboratories obtain completely inconsistent results.

The numbers of ζ, z_A, and z_B values that are greater than 1 are given in Table 1.

For a smaller number of results ($i = 2$ laboratories, $j = 2$ results), the effects are similar.

It appears that there is very high positive correlation between z_A and z_B, particularly for one and two outliers in the case that each laboratory performs tests on three samples and for one outlier in the case of two samples. Pearson product–moment correlation coefficients for z_A versus z_B are given in Table 1.

This good correlation is not profitable. z_B is based on methods of robust location and scale estimation dedicated specifically to small samples [8] and Algorithm A does not concern small samples. The location estimator T_n and scale estimator S_n show monotonicity, in contrast to estimators X^* and s^* obtained using Algorithm A. It turns out that this does not matter for the evaluation of ILC results using parameters such as the z score.

In the present experiment, very large discrepancies between results are introduced. Bias values are 2,…10 times the SD σ and 40,…200 % of mean μ. However, the effectiveness of proposed ζ, z_A, and z_B parameters in detecting incorrect results is very low. The experiment clearly shows the difference between the types of detected discrepancies of test results, which of course results from the nature of the parameter. The ζ parameter is more effective in detecting differences between laboratories, whereas z_A and z_B are better for detecting a laboratory with outliers.

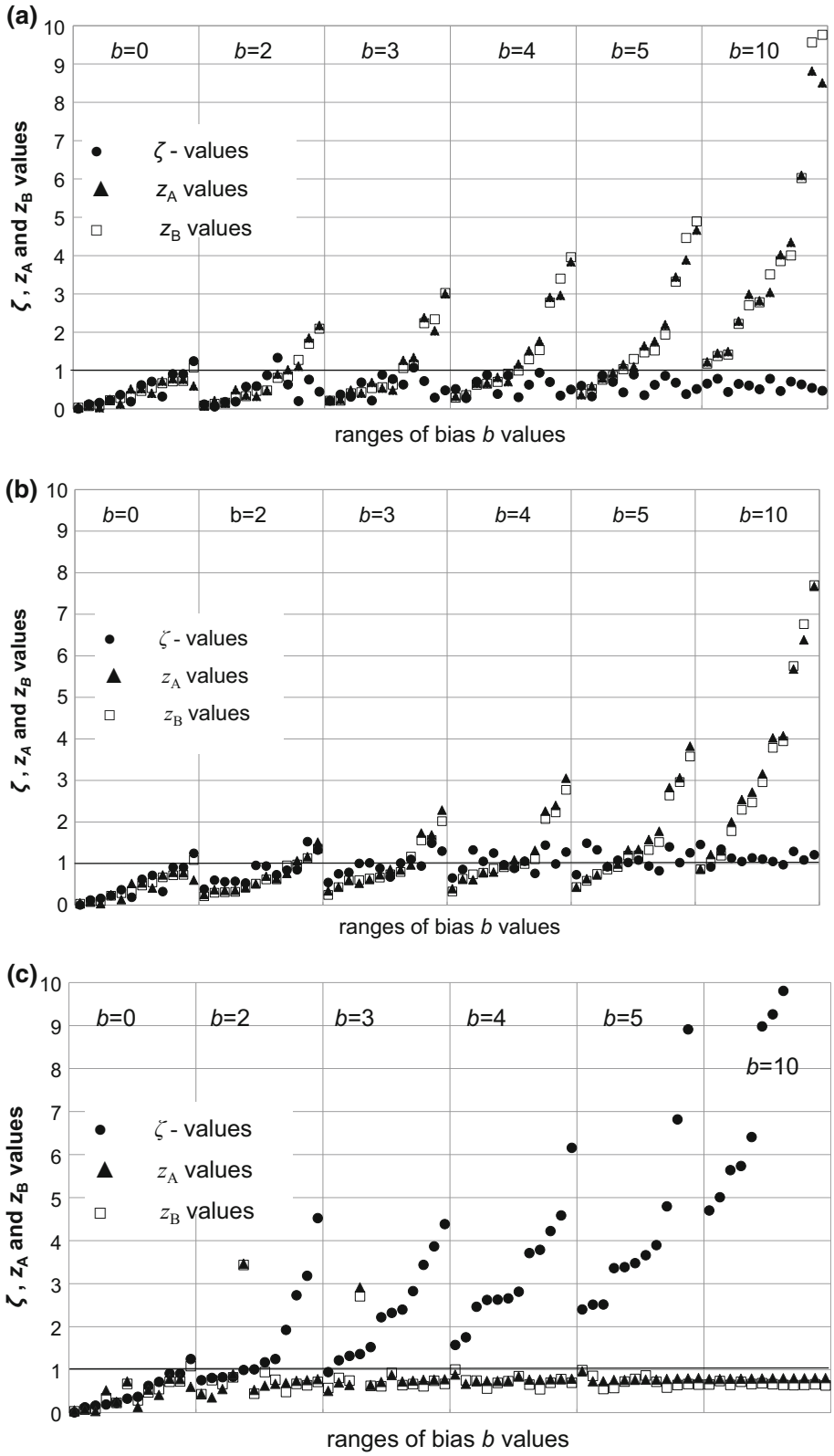

Fig. 1 Dependence of ζ, z_A, and z_B for the second laboratory on the value of bias for **a** one outlier, **b** two outliers, and **c** three outliers introduced for the second laboratory (data within a given range of the b value are arranged in ascending order by z_B for figures **a** and **b** and by ζ for figure **c**, simply for easier visualization.)

Table 1 Effectiveness of the detection of incorrect results, expressed in numbers of ζ, z_A, and z_B values calculated for LAB_2 that are greater than or equal to 1

j Numer of outliers in (in LAB_2),	b Bias value, according to Eq.(8)	i=2 laboratories, j= 3 results				i=2 laboratories, j= 2 results			
		The number of ζ, z_A and z_B values that are greater than or equal 1			Correlation between z_A and z_B values, r [a]	The number of ζ, z_A and z_B values that are greater than or equal 1			Correlation between z_A and z_B values, r [a]
		ζ	z_A	z_B		ζ	z_A	z_B	
0	0	1	1	0		2	2	2	
1	2	1	3	4	0.993	3	3	4	0.999
	3	1	5	5		3	5	5	
	4	0	6	6		2	6	6	
	5	0	8	8		2	9	8	
	10	0	**12**	**12**		1	11	**12**	
2	2	2	2	3	0.995	9	0	0	0.879
	3	6	4	3		11	0	0	
	4	6	4	6		**12**	0	0	
	5	8	7	7		**12**	0	0	
	10	10	11	11		**12**	0	0	
3	2	7	3	3	0.957				
	3	7	1	1					
	4	11	1	1					
	5	**12**	1	0					
	10	**12**	0	0					

Bold values indicate simulations for which effectiveness of the detection of incorrect results was 100 %

[a] Pearson product–moment correlation coefficient

The findings of this experiment are not optimistic, because no statistically reliable parameter for the assessment compliance of results, obtained by two laboratories and for a small number of test results, has been found. Does this mean that such a comparison should not be carried out? In our opinion, such a comparison definitely should be performed. The test method should provide the laboratory customer with confidence that the laboratory has a useful tool for the assessment of the conformity of tested item with specified requirements. Decision making using sample-based location and scale estimators for very small samples is uncertain and may be different for two different laboratories. However, even if no reliable methods of interlaboratory comparison exist, such comparisons give both the laboratory and its client a slightly higher sense of security. Sometimes in such cases, the "researcher's eye" is more useful than statistics. If we take two sets of results, an experienced laboratory worker would immediately find doubtful results.

It is sometimes possible to establish simple criteria for ILC, which are harmonized with criteria for the tested product. There are many possibilities for such criteria. For example, to establish criteria that refer to the suitability of the test method for conformity assessment, one may rely on the permissible product tolerance for the tested product:

$$\frac{U_{SL} - L_{SL}}{\sigma} \leq \kappa, \tag{18}$$

where U_{SL} and L_{SL} are the upper and lower specification limits for the tested item, respectively, and σ is the sample SD for all LAB_1 and LAB_2 results. κ should of course be dependent, as mentioned earlier, on a number of factors and should help to minimize the risk of a different assessment of the tested product at two different laboratories.

Sometimes conformity assessment of a product is based on a value declared by the producer. In such a case, the best solution is to use arbitrarily established criteria based on

experience of the test method and its suitability for conformity assessment. An example of such a criterion is that the SD σ (defined as above) should not be greater than, e.g., 10 % of the test result. As a test result we can use, for example, the robust value X_B calculated in Algorithm B. This idea is based on the maximum permissible variance of the test results, which will allow for a meaningful assessment of the product conformity.

It should be noted that this type of test method most commonly misses data related to precision. Unfortunately, in the process of method development, even by standard committees, exhaustive validation is often lacking, which would be a source of knowledge about the properties of the test method. If it were not so, the data regarding precision (e.g., the SDs of repeatability and reproducibility) could simply be used to establish criteria for the ILC. Even if the assigned value is unknown, knowledge about the precision of the test method presents the possibility of developing a simple criterion based, for example, on values of the repeatability and reproducibility limits published in standards; e.g., the difference between laboratory results should not be greater than the reproducibility limit.

Conclusions

Requirements and rules concerning the organization of proficiency testing or ILC and the analysis of data obtained are not applicable for some kind of tests, when the numbers of laboratories and tests are small and no reference values are available.

It seems to be justified in such a situation to resign actions aimed at ensuring the quality of tests by conducting interlaboratory comparisons and to focus on other aspects, such as the high competence of personnel and the suitability of equipment. However, laboratories, particularly those responsible for carrying out tests of products that affect health and safety, tend to be concerned about the correctness of their test results. An interlaboratory comparison could help them assess whether differences between laboratories are significant and to have more confidence in their results.

The use of statistical methods turns out to have high risk, particularly a high risk of falsely accepting results. The z score parameters z_A and z_B, based on an assigned value, are more effective in detecting a laboratory having outlier results. The ζ parameter, which is based on the difference in results of laboratories and its SDs as described in this article, is better for detecting differences between laboratories. The combination of the two methods (using ζ and z_A or ζ and z_B) can reduce the risk that one of the types of discrepancy is overlooked. However, never do either of these methods or their combination guarantee proper assessment and they should not be used for the main assessment of laboratory performance in such interlaboratory comparisons. It was also shown that methods dedicated to the robust estimation of scale and location in small samples do not improve the efficiency of the "z score"-type parameter in detecting discrepancies of tests results.

In our opinion, the best option is to use arbitrarily defined criteria based on the experience of laboratories, suitable for the requirements of the tested product, the aim of the tests, and other known risk factors.

Simultaneously to this work (unexpectedly for authors of the paper), new version of ISO 13528 [17] has been published. In informative Annex D1 some conclusions on procedures for small numbers of participants has been shown. The external criteria independent of the participants' results are preferred in ISO for small number of participants. Also unreliability of some procedures used for the performance evaluation for too small number of participants has been underlined in the standard. Thus, our conclusions are consistent with information given in the new standard.

Assessment of the reliability of small populations of test results is a difficult but necessary problem to solve in terms of not only ILC but also the conformity assessment of tested product and will be the subject of further work of the authors of this article.

Appendix 1: code for calculation of location and scale estimators according to Algorithm B.

(MATLAB language, MATLAB R2014a (8.3.0.532) by MathWorks, Inc.)

```
function [location,iter,sigma]=assignedvalue(Y)
X=reshape(Y,1,[]); % converting data to a vector

f =inline('(exp(x)-1).*((exp(x)+1).^(-1))','x');
fp=.4132; % constant used instead of a derivative of the f function for better sta-
bility of the Newton-Rhapson algorithm

N = 10000; % maximal allowed number of iterations
eps = 1.e-6; % maximal allowed error
maxval = max(X); % value used to state that the algorithm does not converge
xn=median(X); % initial value that is adjusted during iteration
sigma=1.4826*median(abs(X-xn));
if length(X)<4
    location = xn;  % version for very small samples (size 2,3)
    iter = 0;
        return;

    elseif length(X)==4
        sigma=sigma*1.09545;     % removing bias for particular sizes of samples
    elseif length(X)==5
        sigma=sigma*1.06904;
    elseif length(X)==6
        sigma=sigma*1.05409;
    elseif length(X)==7
        sigma=sigma*1.04447;
    elseif length(X)==8
        sigma=sigma*1.03775;
    elseif length(X)==9
        sigma=sigma*1.03280;
    elseif   length(X)==10
        sigma=sigma*1.02899;  % for size of sample n > 10, constants could be ignored
                              % because they are close to 1
end

if sigma < xn/1000
    iter = 0;
    location = xn;
    return;
end

while (N>0)
  xn = xn+sigma*mean(f((X-xn)/sigma))/fp;
  if abs(mean(f((X-xn)/sigma)))<eps
    location=xn;iter=10000-N;
    return;
  end;

  if abs(mean(f((X-xn)/sigma)))>maxval
    disp(['iterations = ',num2str(iter)]);
    error('Solution diverges');
    break;
  end;
  N = N - 1;
  location = xn;
end;
error('Solution diverges');
return;
```

References

1. ISO, IEC 17025 (2005) General requirements for the competence of testing and calibration laboratories. International Organization for Standardization/International Electrotechnical Commission, Geneva
2. EA-4/18 (2010) Guidance on the level and frequency of proficiency testing participation. http://www.european-accreditation.org/publication/ea-4-18-inf-rev00-june-2010
3. ISO, IEC 17043 (2010) Conformity assessment—general requirements for proficiency testing. International Organization for Standardization/International Electrotechnical Commission, Geneva
4. ISO 13528 (2005) Statistical methods for use In proficiency testing by interlaboratory comparisons. International Organization for Standardization, Geneva
5. Kuselman I, Fajgelj A (2010) IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for limited number of participants- chemical analytical laboratories (IUPAC Technical Report). Pure Appl Chem 82(5):1099–1135
6. Regulation (EU) No 305/2011 Of The European Parliament And Of The Council of 9 March 2011 laying down harmonised conditions for the marketing of construction products and repealing Council Directive 89/106/EEC
7. ISO 5725–5 (1998) Accuracy (trueness and precision) of measurement methods and results—Part 5: alternative methods for the determination of the precision of a standard measurement method. International Organization for Standardization, Geneva
8. Rousseeuw PJ, Verboven S (2002) Robust estimation in very small samples. Comput Stat Data Anal 40:741–758
9. Kacker RN, Forbes AB, Kessel R, Sommer K-D (2008) Bayesian posterior predictive p-value of statistical consistency in interlaboratory evaluations. Metrologia 45:512–523
10. Kacker RN, Forbes A, Kessel R, Sommer K-D (2008) Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluation. Metrologia 45:257–264
11. Kacker RN, Forbes A, Kessel R, Sommer K-D (2010) Assessing differences between results determined according to the guide to the expression of uncertainty in measurement. J Res Natl Inst Stand Technol 115:453–459
12. Kessel R, Kacker RN (2011) Combining results from multiple evaluations of the same measurand. J Res Natl Inst Stand Technol 116:809–820
13. JCGM 100 (2008) Evaluation of measurement data: guide to the expression of uncertainty in measurement (GUM). http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf
14. JCGM 200 (2012) International vocabulary of metrology: basic and general concepts and associated terms (VIM), 3rd edn http://www.bipm.org/vim
15. Richard M, Brugger A (1969) Note on unbiased estimation of the standard deviation. Am Stat 23(4):32
16. Huang H (2015) Optimal estimator for uncertainty-based measurement quality control. Accred Qual Assur 20:97–106
17. ISO 13528 (2015) Statistical methods for use In proficiency testing by interlaboratory comparisons. International Organization for Standardization, Geneva