

Ilya Kuselman

Comparability of analytical results obtained in proficiency testing based on a metrological approach

Received: 4 September 2005
Accepted: 9 October 2005
Published online: 16 December 2005
© Springer-Verlag 2005

I. Kuselman (✉)
The National Physical Laboratory of
Israel (INPL),
Danciger “A” Bldg, Givat Ram,
Jerusalem 91904, Israel
e-mail: ilya.kuselman@moital.gov.il

Abstract A “yes–no” type of criterion is proposed for the assessment of comparability of proficiency testing (PT) results when the PT scheme is based on a metrological approach, i.e. on the use of a reference material as the test sample, etc. The criterion tests a null hypothesis concerning the insignificance of a bias of the mean of the results from a traceable value certified in the reference material used for the PT. Reliability of such assessment is determined by the probabilities of not rejecting the null hypothesis when it is true, and

rejecting it when it is false (the alternative hypothesis is true). It is shown that a number of chemical, metrological and statistical reasons should be taken into account for careful formulation of the hypotheses, enabling the avoidance of an erroneous assessment of the comparability. The criterion can be helpful for PT providers and laboratory accreditation bodies in analysis of PT results.

Keywords Comparability · Traceability · Proficiency testing · Test of hypotheses · Reliability

Introduction

The concept of comparability (equivalence) of measurement results—“tested once, accepted everywhere” [1]—is increasingly important since it allows to minimize technical barriers in trade, to improve environmental monitoring and medical care and to cut down expenditures for international cooperation. While results obtained under repeatability conditions can be compared directly, results obtained by different laboratories in different countries and at different times are comparable through their relationship to the same reference which is an internationally agreed and recognized measurement standard. This strategy of linking results to a reference by an unbroken chain of comparisons or calibrations is termed “traceability”. Traceability of a measurement result is one of the principal tools required for comparability [1]. It is underscored by adoption of standards ISO 17025 for calibration and testing laboratories [2], ISO 15195 for reference measurement laboratories in medicine [3] and of GLP-GMP standards in pharmaceutical and other industries [4, 5]. To design the traceability chains and understand the degree of equivalence of measurement results (their comparability), measurement uncertainty quantification is necessary. The quantification procedure starts from

a careful measurand/analyte definition, includes analysis of all the measurement steps and the traceability chains as sources of uncertainty components and takes into account the method validation data [6].

Practical estimation of comparability of measurement/testing/analytical results is based on interlaboratory comparisons (intercomparisons). Key comparisons conducted on the best measurement capability level of national metrology institutes are organized by CCQM (Consultative Committee for Amount of Substance: Metrology in Chemistry) at the International Bureau of Weights and Measures (BIPM) within the framework of the Convention of the Meter. Results of these comparisons are published in the key comparison data base (KCDB) on the BIPM webpage: www.bipm.org, and are widely discussed in the literature [7–9].

Intercomparisons organized on the routine measurement level of field laboratories are named “proficiency testing” (PT) [10] since they are used mostly for assessment of a laboratory performance. There are international documents regulating requirements for the competence of providers of PT schemes [10, 11], statistical methods to be used in PT by intercomparisons [12], organization of PT schemes and their implementation for analytical (chemical) laboratories

[13] and other. PT is a topic of numerous international conferences and workshops [14]. One can find the necessary PT scheme announced by PT providers in the international database EPTIS [15], in *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement* (Springer) and in the Internet. The providers are metrology institutes, accreditation bodies and private companies, since PT is nowadays also a business affected by ISO 17025 [2] and other standards for laboratory accreditation, which define its performance as an important element of its technical competence.

Two main steps are common for all PT schemes: (1) estimation of the assigned value of analyte concentration in the test items/reference material and quantification of the value's uncertainty, including components arising from the material homogeneity and stability and (2) calculation of performance statistics and assessment of the laboratory performance [10]. As test items, portions of a certified reference material (CRM), in the ideal case, are distributed among the laboratories participating in the PT. If CRM is expensive or not available, an in-house reference material (IHRM) or a spike with the traceable property value can be helpful [16, 17]. It is a metrological approach to proficiency testing [18]. The mean or the median of PT results is also used as the assigned/consensus value of the material sent to the laboratories participating in the PT [13]. Performance of a laboratory/PT participant is assessed by the difference between the laboratory result and the assigned value, in comparison to the sum of their uncertainties, or to the standard deviation of the PT results, or to a target standard deviation that can be used as a parameter of the analysis fitness for purpose [12, 18, 19].

Since the laboratory performance is assessed individually for each PT participant, even in the case when the performance of the majority of them is found to be successful, the question about comparability of the PT results (as a whole) still remains unsolved. The purpose of the present paper is to develop a criterion allowing a PT provider and/or an accreditation body to assess the comparability of results obtained for a laboratory proficiency testing based on a metrological approach.

Hypotheses and the criterion for their testing

One can imagine a situation when CRM, IHRM or a spike, with a traceable certified/assigned value of the analyte concentration C_{cert} and standard uncertainty σ_{cert} , is used for proficiency testing of a (theoretically) infinite population of laboratories that produced results with the mean C_{PT} and standard deviation σ_{PT} . For the sake of simplicity, the data of reference material certification and PT results are considered as independent random events having both normal distributions with parameters C_{cert} , σ_{cert} and with C_{PT} , σ_{PT} , correspondingly.

To formulate a "yes-no" type of criterion for assessment of PT results comparability, the null hypothesis H_0 concerning insignificance of the bias $|C_{\text{PT}} - C_{\text{cert}}|$ has to be established. For example, H_0 may consist of the assumption

that the bias exceeds the value σ_{cert} (approximately equal to half the expanded uncertainty of C_{cert} given in the certificate of the reference material) by a value which is insignificant in comparison with random interlaboratory errors of the analysis. In this case, the null hypothesis has the following form:

$$H_0 : |C_{\text{PT}} - C_{\text{cert}}| \leq [(0.3 \sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2}, \quad (1)$$

The alternative hypothesis H_1 assumes that the bias exceeds σ_{cert} significantly, e.g. by analogy with a similar situation analyzed in [20],

$$H_{11} : |C_{\text{PT}} - C_{\text{cert}}| = 2.0[(0.3 \sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2}, \quad (2)$$

$$H_{12} : |C_{\text{PT}} - C_{\text{cert}}| = 2.1[(0.3 \sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2}, \quad (3)$$

etc.

In practice, laboratories participating in PT form a statistical sample (from the population) of size N , i.e. only N laboratories are sending their results. Therefore, the criterion for not rejecting H_0 is

$$|C_{\text{PT/av}} - C_{\text{cert}}| + t_{1-\alpha/2} S_{\text{PT}} / N^{1/2} \leq [(0.3 \sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2}, \quad (4)$$

where $C_{\text{PT/av}}$ and S_{PT} are the sample estimates of C_{PT} and σ_{PT} calculated from the same N results as the sample average and standard deviation, correspondingly; the left-hand side of the expression represents the upper limit of the confidence interval for the bias $|C_{\text{PT}} - C_{\text{cert}}|$; $t_{1-\alpha/2}$ is the quantile of the Student's distribution for the number of degrees of freedom $N-1$; the value $1-\alpha/2$ is the probability of the bias $|C_{\text{PT}} - C_{\text{cert}}|$ not exceeding the upper limit of its confidence interval. Substituting the ratio $\sigma_{\text{cert}}/\sigma_{\text{PT}} = \gamma$ into Eq. (4), the following expression can be obtained:

$$|C_{\text{PT/av}} - C_{\text{cert}}| / S_{\text{PT}} \leq (\sigma_{\text{PT}} / S_{\text{PT}}) (0.09 + \gamma^2)^{1/2} - t_{1-\alpha/2} / N^{1/2}. \quad (5)$$

Since $S_{\text{PT}}/\sigma_{\text{PT}} = [\chi_{\alpha/2}^2/(N-1)]^{1/2}$, where χ^2 is the quantile of chi-square distribution for the number of degrees of freedom $N-1$, the criterion is transformed again:

$$|C_{\text{PT/av}} - C_{\text{cert}}| / S_{\text{PT}} \leq \{[(N-1)/\chi_{\alpha/2}^2] (0.09 + \gamma^2)\}^{1/2} - t_{1-\alpha/2} / N^{1/2}. \quad (6)$$

Table 1 Comparability assessment according to criterion (6): the bias norms in S_{PT} units

γ	N						
	5	10	15	20	30	40	50
0.4	0.20	0.20	0.23	0.26	0.30	0.32	0.34
0.7	0.95	0.68	0.65	0.64	0.65	0.66	0.67
1.0	1.76	1.19	1.09	1.06	1.03	1.02	1.02

Table 2 PT results of aluminium determination in SRM 2690 (simulated in % by weight)

Lab. No	Result (%)	Lab. No	Result (%)
1	12.76	16	12.60
2	12.19	17	12.81
3	12.68	18	12.39
4	12.21	19	11.96
5	12.96	20	11.91
6	12.27	21	11.86
7	11.96	22	12.32
8	12.03	23	12.53
9	11.88	24	12.84
10	11.97	25	12.67
11	12.23	26	12.86
12	12.48	27	12.75
13	12.69	28	12.66
14	12.21	29	11.99
15	11.98	30	12.61
$C_{PT/av}$	12.30	$C_{PT/av}$	12.38
S_{PT}	0.34	S_{PT}	0.35

Table 1 gives the numerical values for the right-hand side of the criterion at $\alpha=0.05$. These values are the norms for the bias of the average PT result from the analyte concentration certified in the reference material (in S_{PT} units). The value γ is set based on requirements of the analysis, taking into account a σ_{PT} value equal to the standard analytical/measurement uncertainty or to the target standard deviation (calculated using the Horwitz curve [12] or another database).

For example, if standard reference material SRM 2690 (NIST, USA) is chosen for PT of aluminium determination in coal fly ashes, $C_{cert}=12.35\%$ by weight and $\sigma_{cert}=0.28/2=0.14\%$ ($\pm 0.28\%$ is the expanded uncertainty shown in the certificate). According to the ASTM standard [21], the means of the results of duplicate determinations carried out by different laboratories on riffled splits of the analysis sample should not differ by more than 2.0% for Al_2O_3 , i.e. 1.06% for aluminium. Since the range for two laboratory results is limited by the standard, $\sigma_{PT}=1.06/2.77=0.38\%$, where 2.77 is the 95% percentile of the range distribution [22], the value $\gamma=0.14/0.38=0.4$. Simulated statistical samples of the PT results are given in Table 2. Comparability of results of the first 15 laboratories can be assessed as satisfactory by the norm in Table 1 for $\gamma=0.4$ (0.23), since $|C_{PT/av}-C_{cert}|=|12.30-12.35|=0.05 < 0.23S_{PT}=0.23 \times 0.34=0.08\%$ by weight. The same is true concerning comparability of results of all the 30 laboratories (the norm is 0.30): $|C_{PT/av}-C_{cert}|=12.38-12.35|=0.03 < 0.30S_{PT}=0.30 \times 0.35=0.11\%$.

Reliability of the assessment

Reliability in such comparability assessment is determined by the probabilities of not rejecting the null hypothesis H_0 when it is true, and rejecting it when it is false (the

alternative hypothesis H_1 is true). Criterion (6) does not allow for the rejecting of hypothesis H_0 for the probability $1-\alpha/2$, when it is true. Probability of an error of type I by this criterion (to reject the H_0 hypothesis when it is true) is $\alpha/2$. The probability of rejecting H_0 , when it is false, i.e. when the alternative hypothesis H_1 is actually true (the criterion power [23]) is:

$$P = \phi\{(t_{\alpha/2} + \lambda)/[1 + t_{1-\alpha/2}^2/2(N-1)]^{1/2}\}, \quad (7)$$

where ϕ is the function of the normalized normal distribution, and

$$\lambda = [|C_{PT}-C_{cert}|-\sigma_{PT}(0.09 + \gamma^2)^{1/2}]/(\sigma_{PT}/N^{1/2}). \quad (8)$$

The value of the deviation parameter λ is calculated by substituting the expression $|C_{PT}-C_{cert}|$ in Eq. (8) for its value, corresponding to the alternative hypothesis. For example, for hypothesis H_{11} in Eq. (2) this is $2.0[(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}$. Therefore, for H_{11} , the parameter is $\lambda=[(0.09+\gamma^2)N]^{1/2}$, and for H_{12} in Eq. (3) it is $\lambda=1.1[(0.09+\gamma^2)N]^{1/2}$. The probability of an error of type II (not rejecting the H_0 when it is false) equals $\beta=1-P$. Both operational characteristics of the criterion P and β are shown in Fig. 1 at $\alpha=0.05$ for the alternative hypotheses H_{11} and H_{12} , and for different γ values and different numbers N of the PT participants.

For example, reliability of the assessment using the hypotheses H_0 against H_{11} for the above-described PT scheme for aluminium determination in coal fly ashes ($\gamma=0.4$), can be characterized by 1) probability $1-\alpha/2=0.975$ for correctly assessing comparability as successful (not rejecting the null hypothesis H_0 when it is true) for any number N of the PT participants, and by 2) probability $P=0.42$ for correctly assessing comparability as unsuccessful for $N=15$ PT results (rejecting H_0 when the alternative hypothesis

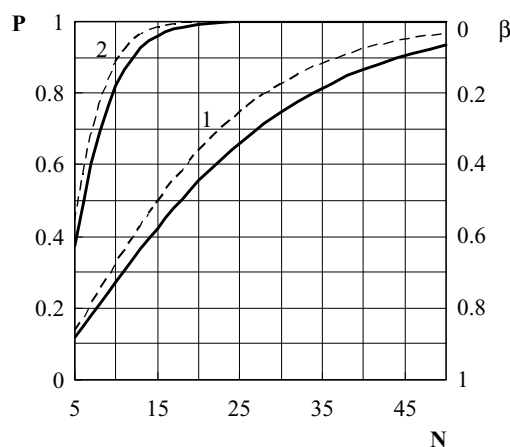


Fig. 1 Power P of the criterion and probability β of an error of type II (dependent on the number N of laboratories participating in PT) for probability $\alpha/2=0.025$ of an error of type I. *Solid lines* are for the case when the null hypothesis H_0 is tested against the alternative hypotheses H_{11} , and *dotted lines* are when it is tested against the alternative hypothesis H_{22} . Curves 1 correspond to $\gamma=0.4$, and curves 2 correspond to $\gamma=1.0$

H_{11} is true), and probability $P=0.75$ of the same event for $N=30$ results. Probability $\alpha/2$ of an error of type I is 0.025 for any N , while probability β of a type II error is 0.58 for $N=15$, and 0.25 for $N=30$, etc.

Discussion

The operational characteristics of the criterion expressed in Eq. (6) and shown in Fig. 1 are high enough ($P>0.5$) for the number of PT participants $N\geq 20$. However, it is known that reliability of the assessment and probabilities of erroneous decisions based on such criterion depend on how the hypotheses have been formulated, as well as on the criterion for their testing [20]. In particular, the criterion power increases with λ values from the alternative hypothesis H_{11} to the hypothesis H_{12} , since the last one is “further” from the null hypothesis H_0 . Thus, detecting greater bias is simpler: for H_0 it is $[(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}$, for H_{11} – $2.0 \times [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}$ and for H_{12} – $2.1 \times [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}$.

The role of the ratio γ of the population standard deviations σ_{cert} and σ_{PT} is more complicated. Formally the larger γ is, the larger λ and the criterion power are. However, $\gamma>1$ ($\sigma_{cert}>\sigma_{PT}$) makes no metrological sense even when IHRM or a spike is used, since it means that either less accurate methods were applied for the material preparation and study than those used by laboratories participating in the PT, or the material is not suitable because of its inhomogeneity and/or instability. On the other hand, if it is possible to neglect the uncertainty of the reference material certification, as compared with random interlaboratory errors of the analysis, i.e. when $\sigma_{cert}\ll\sigma_{PT}$, then the value $\gamma<0.3$ can be considered as insignificant. In such cases, Eq. (6) is reduced to a simpler one, similar to that described by Kuselman and Skryabina [24], since the null hypothesis by Eq. (1) is transformed here into a different one based on the statement that the bias of PT results from the certified value of the reference material is also negligible in comparison with random interlaboratory errors: $|C_{PT}-C_{cert}|\leq 0.3\sigma_{PT}$. The corresponding criterion is less powerful and requires $N=50-100$ for achieving the operation characteristics demonstrated for criterion (6) in Fig. 1. When both σ_{cert} and σ_{PT} are to be taken into account as values of the same order, i.e. when $0.4\leq\gamma\leq 1$, the null hypothesis can be formulated as $|C_{PT}-C_{cert}|\leq 0.3(\sigma_{PT}^2 + \sigma_{cert}^2)^{1/2}$, i.e. the bias is negligible in comparison with the sum of random interlaboratory errors and the material certification uncertainty. This hypothesis leads to the criterion being stricter than for Eq. (6), which is probably more appropriate for comparability assessment of key comparison results.

It should also be taken into consideration that information about the distributions (of both analytical results used for

the reference material certification and of the PT results) is limited by the small sizes of statistical samples of experimental data. Therefore, the comparability assessment can be adequate only as long as the postulate concerning the normal distributions of the results fits the reality. In those cases where the distributions differ from the normal ones, and it is impossible to determine and to remove the causes of that difference, similar criteria can be formulated by non-parametric statistics methods; for example, using the sign test of hypotheses concerning the bias value $|M_{PT}-C_{cert}|$, where M_{PT} is the median of the population of PT results at the hypothetically infinite number of participants. Thus, the null hypothesis by Eq. (1) is transformed here into $H_0 : |M_{PT}-C_{cert}|\leq [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}$ and so on [25]. However, the power efficiency of the sign test in relation to the t -test (ratio of the sizes N of statistical samples from normal populations allowing the same power) is from 0.96 for $N=5$ to 0.64 for infinite N [26]. Therefore, the power of a non-parametric criterion will always be less than the power of a parametric one for the same number N of PT participants.

From the chemical/metrological point of view, discussion of the results comparability should be restricted by the definition of the analyte and the matrix for which C_{cert} and σ_{cert} are quantified. Therefore, the adequacy of the reference material used is very important [27]. On the other hand, if the reference material prepared for PT is not certified and a consensus value (the average or the median of the PT results) is used instead of C_{cert} , traceability of this value is questionable and comparability of the PT results cannot be assessed in the sense of “tested once, accepted everywhere”. In such cases, especially when number N of the participants is limited, a local comparability, i.e. among the participants only, is tested based on known requirements to the method reproducibility [28].

Conclusions

1. Comparability of measurement results in a PT scheme can be assessed based on a “yes–no” type of criterion for testing the null hypothesis regarding the insignificance of the bias of the results mean from the traceable value certified in the reference material used for the PT. The reliability of such assessment is determined by the probability of not rejecting the null hypothesis when it is true, and rejecting it when it is false and the alternative hypothesis is true.
2. A combination of chemical/metrological and statistical knowledge is necessary for careful formulation of the null hypothesis, since different hypotheses can lead to different decisions about comparability of the measurement results obtained in the same PT scheme.

References

1. EURACHEM/CITAC Guide (2003) Traceability in chemical measurement; a guide to achieving comparable results in chemical measurement, EURACHEM/CITAC, www.eurachem.ul.pt
2. ISO (1999) General requirements for the competence of testing and calibration laboratories, ISO/IEC Standard 17025, ISO, Geneva
3. ISO (2003) Laboratory medicine: requirements for reference measurement laboratories, BS EN ISO Standard 1519, ISO, Geneva
4. Garner WY, Barge MS, Ussary JP (eds) (1992) Good laboratory practice standards, ACS Professional Reference Book, ACS, Washington DC
5. Miller JM, Crowther JB (eds) (2000) Analytical chemistry in a GMP environment. A practical guide. Wiley, New York
6. EURACHEM/CITAC Guide (2000) Quantifying uncertainty in analytical measurement, 2nd edn., EURACHEM/CITAC, www.eurachem.ul.pt
7. Quinn T (2004) *Accred Qual Assur* 9:533–538
8. Kaarls R (2004) *Accred Qual Assur* 9:530–532
9. Kaarls R (2005) *CITAC News* 2005:4–6
10. ISO/IEC Guide 43 (1997) Proficiency testing by interlaboratory comparisons. Part 1: Development and operation of proficiency testing schemes. Part 2: Selection and use of proficiency testing schemes by laboratory accreditation bodies
11. ILAC (2000) Guidelines for the requirements for the competence of providers of proficiency testing schemes, ILAC G13, ILAC, Rhodes, Australia
12. ISO (2002) Statistical methods for use in proficiency testing by interlaboratory comparisons, ISO/DIS Standard 13528 (provisional version), ISO, Geneva
13. ISO/AOACI/IUPAC (2004) The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories, ISO, Geneva
14. Ornemark U (2005) *CITAC News* 2005:20
15. EPTIS (2005) The International Proficiency Testing Information System EPTIS: a PT directory for Europe and the Americas <http://www.eptis.bam.de>
16. Kuselman I, Pavlichenko M (2004) *Accred Qual Assur* 9:399–403
17. Kuselman I (2004) *J. of Metrology Society of India* 19/4:245–252
18. Armishaw P, King B, Millar RG (2003) *Accred Qual Assur* 8:184–190
19. Kuselman I, Papadakis I, Wegscheider W (2001) *Accred Qual Assur* 6:78–79
20. Kuselman I (1993) *Talanta* 40/1:1–7
21. ASTM (1974) Standard methods of analysis of coal and coke ash, ASTM Standard D 2795-74, ASTM, <http://www.astm.org>
22. Owen DB (1962) Handbook of statistical tables, Addison-Wesley, London, pp.138–139
23. Miller JC, Miller JN (1993) Statistics for analytical chemistry, 3rd edn. Horwood, Bodmin, England
24. Kuselman I, Skryabina LG (1987) *Indust Lab* 53/5:444–447
25. Kuselman I, Skryabina LG (1983) *Metrologia* (in Russian) 3:59–64
26. Dixon WJ, Massey FJ Jr. (1969) Introduction to statistical analysis, 3rd edn., McGraw-Hill, Tokyo, p. 340
27. Kuselman I (2004) *Accred Qual Assur* 9:591–596
28. Kardash-Strochkova E, Tur'yan YaI, Kuselman I, Brodsky N (2002) *Accred Qual Assur* 7:250–254