Accred Qual Assur (2007) 12:391–398
DOI 10.1007/s00769-006-0247-0

GENERAL PAPER

Implementation of proficiency testing schemes for a limited number of participants

Maria Belli · Stephen L. R. Ellison · Ales Fajgelj · Ilya Kuselman · Umberto Sansone · Wolfhard Wegscheider

Received: 4 September 2006 / Accepted: 19 December 2006 / Published online: 21 February 2007
© Springer-Verlag 2007

Abstract A metrological background for the selection and use of proficiency testing (PT) schemes for a limited number N of laboratories-participants (less than 20–30) is discussed. The following basic scenarios are taken into account: (1) adequate matrix certified reference materials (CRM) or in-house reference materials (IHRM) with traceable property values are available for PT use as test items; (2) no appropriate matrix CRM is available, but a CRM or IHRM with traceable property values can be applied as a spike or similar; (3) only an IHRM with limited traceability is available. The discussion also considers the effect of a limited population of PT participants N_p on statistical assessment of the PT results for a given sample of N responses from this population. When N_p is finite and the sample fraction N/N_p is not negligible, a correction to the statistical parameters may be necessary. Scores suitable for laboratory performance assessment in such PT schemes are compared.

Keywords Proficiency testing · Sample size · Population · Traceability · Measurement uncertainty

Presented at the 3rd International Conference on Metrology, November 2006, Tel Aviv, Israel.

M. Belli
National Agency for Environmental Protection and Technical Services (APAT),
Via di Castel Romano, 100, 00128 Rome, Italy

S. L. R. Ellison
LGC Limited, Queens Road, Teddington,
Middlesex TW11 0LY, UK

A. Fajgelj · U. Sansone
Agency's Laboratories,
International Atomic Energy Agency (IAEA),
Seibersdorf, Wagramer Strasse 5,
1400 Vienna, Austria

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

W. Wegscheider
University of Leoben, Franz-Josef Strasse 18,
8700 Leoben, Austria

Introduction

The International Harmonized Protocol for the proficiency testing (PT) of analytical chemistry laboratories adopted by the International Union of Pure and Applied Chemistry (IUPAC) in 1993 [1] has been revised in 2006 [2]. Statistical methods for use in PT [3] have been published as a complementary standard to ISO/IEC Guide 43, which describes PT schemes based on interlaboratory comparisons [4]. There are International Laboratory Accreditation Cooperation (ILAC) guidelines defining requirements for the competence of PT providers [5]. Guidelines for PT application in specific sectors, such as clinical laboratories [6] (recently revised), have also been widely available; in some other sectors they are under development.

These documents are, however, oriented mostly towards PT schemes for a relatively large number N of laboratories-participants (more than 20–30), referred to from here onwards as "large schemes." This is important from a statistical point of view, since most statistical methods used in PT become increasingly unreliable with $N<30$, especially for $N<20$. For exam-

ple, uncertainties in estimates of location (such as mean and median) are sufficiently small to neglect in scoring as N increases to approximately 30, but they cannot be safely neglected with $N<20$. Non-normal distributions are harder to identify for small N. Robust statistics, too, are not usually recommended for $N<20$. Although the methods continue to down-weight very extreme outliers successfully and still perform at least as well as, or better than, the mean, they do become less reliable in coping with the increasingly sporadic appearance of a small proportion of more modest outlying values [7]. Therefore, the certified/assigned value of the PT test material C_{cert} cannot be safely calculated from the PT results as a consensus value: its uncertainty becomes large enough to affect scores in ''small schemes,'' that is, schemes with small numbers of participants ($N<20$–30).

Moreover, if the size N_p of the population of laboratories participating in PT is not infinite, and the size of the statistical sample N is more than 5–10% of N_p, the value of the sample fraction $\rho=N/N_p$ may need to be taken into account [8].

Thus, the implementation of small PT schemes is not a routine task. Such schemes are quite often required for the analysis of materials and/or for environmental analysis specific to a local region, for an industry under development, for the analysis of unstable analytes, for a local laboratory accreditation body to control the performance of less numerous accredited laboratories, etc. Therefore, the IUPAC Interdivisional Working Party on Harmonization of Quality Assurance started a new project [9], with the aim of developing guidelines which could be helpful for PT providers and accreditation bodies in the solution of this task.

A metrological background for the guidelines, including possible criteria for the implementation of small PT schemes, is discussed in the present position paper.

Approach

The difference between the population parameters and the corresponding sample estimates increases with decreasing sample size N. In particular, a sample mean $c_{PT/avg}$ of N PT results can differ from the population mean C_{PT} by up to $\pm 1.96\sigma_{PT}/\sqrt{N}$ with 95% probability, 1.96 being the appropriate percentile of the normal distribution for a two-sided 95% interval and σ_{PT} is the population standard deviation of the results. Dependence of the upper limit of the interval for the expected Bias$=|c_{PT/avg}-C_{PT}|$ on N is shown (in units of σ_{PT}) in Fig. 1, where the range N=20–30 is marked out by the gray bar. Even for N=30, the bias may reach $0.36\sigma_{PT}$ at the 95% level of confidence.

Similarly, the sample standard deviation s_{PT} is expected to be in the range $\sigma_{PT}[\chi^2\{0.025, N-1\}/(N-1)]^{1/2} \le s_{PT} \le \sigma_{PT}[\chi^2\{0.975, N-1\}/(N-1)]^{1/2}$ with 95% probability, where $\chi^2\{\alpha, N-1\}$ is the 100α percentile of the χ^2 distribution at $N-1$ degrees of freedom. The dependence of the range limits for s_{PT} on N is shown in Fig. 2 (again, in σ_{PT} units), also with the range N=20–30 marked by the gray bar. For example, for N=30 the upper 95% limit for s_{PT} is $1.26\sigma_{PT}$. In other words, s_{PT} can differ from σ_{PT} for N=30 by over 25% rel. at the level of confidence of 0.95. For $N<30$ the difference

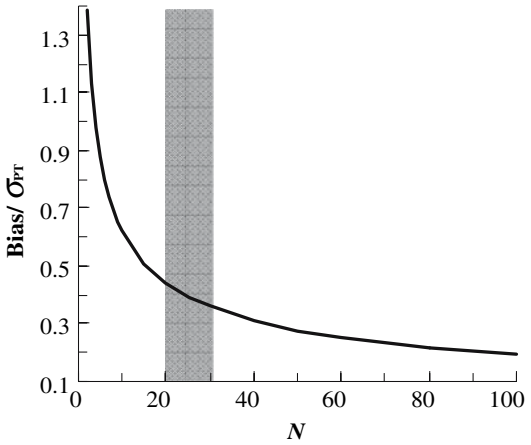

Fig. 1 Dependence of the sample mean bias on the number N of PT results, in units of σ_{PT}. The *line* is the upper 97.5th percentile, corresponding to the upper limit of the two-sided 95% interval for the expected bias. The range of N=20–30, intermediate between small and large sample sizes, is shown by the *gray bar*

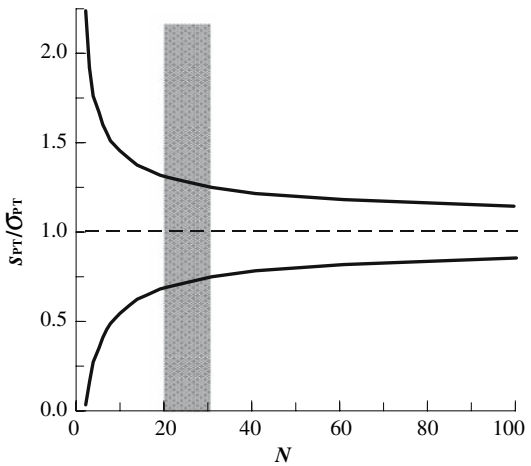

Fig. 2 Dependence of the sample standard deviation s_{PT} on the number N of PT results, in units of σ_{PT}. The *solid* lines show the 2.5th (*lower line*) and 97.5th (*upper line*) percentiles for s_{PT}. The *dashed* line is at $s_{PT}/\sigma_{PT}=1.0$ for reference. The *gray bar* shows the range of intermediate sample sizes (N=20–30)

between the sample and the population characteristics increases with decreasing N, and especially dramatically for the standard deviation when $N<20$.

While consensus mean values are less affected than observed standard deviations, uncertainties in the consensus means are relatively large in small schemes, and will practically never meet the guidelines for unqualified scoring suggested in the IUPAC Harmonized Protocol [2]. It follows that the scoring for small schemes should usually avoid simple consensus values. Methods for obtaining traceable assigned values $C_{cert}$are to be used wherever possible to provide comparable PT results [10].

The high variability of dispersion estimates in small statistical samples has special implications for scoring based on observed participant standard deviation s_{PT}. This practice is already recommended against, even for large schemes [3], on the grounds that it does not provide for the consistent interpretation of scores from one round (or scheme) to the next. For small schemes, the variability of s_{PT} magnifies the problem. It follows that scores based on the observed participant standard deviation should not be applied in such a case. If a PT provider can set a normative population standard deviation σ_p on fitness-for-purpose grounds, z-scores, which compare a result bias from the assigned value with σ_p, can be calculated in a small scheme in the same manner as recommended in [1–4] for a large scheme. The condition is only that the standard uncertainty of the assigned value u_{cert} is insignificant in comparison to σ_p ($u_{cert}^2<0.1\sigma_p^2$). When information necessary to set σ_p is not available, and/or u_{cert} is not negligible, the information included in the measurement uncertainty $u(x_i)$ of the result x_i reported by the ith laboratory is helpful for performance assessment using *zeta*-scores and/or E_n numbers [2, 3]. Such assessment is problematic when participants have a poor understanding of their uncertainty [3]. However, uncertainty data are increasingly required by customers of laboratories, and laboratories should, accordingly, be checking their uncertainty evaluation procedures [2], especially those laboratories that claim compliance with the ISO 17025 standard [11]. Incorporating the uncertainty information provided by laboratories into the interpretation of PT results can play a major role in improving their understanding of this subject [3]. It may also be important for a small scheme that laboratories working according to their own fitness-for-purpose criteria (for example, in conditions of competition) can be judged by individual criteria based on their declared uncertainty values [2].

The approach based on the traceability of assigned values of test items providing the comparability of PT results, and on scoring PT results taking into account uncertainties of the assigned values and uncertainties of the results, has been described as a ''metrological approach'' [12–15].

Two main steps are common for any PT scheme using this approach: (1) establishment of the traceable assigned value, C_{cert}, of analyte concentration in the test items/reference material and quantification of the value's standard uncertainty u_{cert}, including components arising from the material homogeneity and stability during the PT round; and (2) the calculation of fitness-for-purpose performance statistics and assessment of the laboratory performance. For the second step, it may additionally be necessary to take into account the small population size of laboratories able to take part in the PT. These issues are considered below.

Value assignment

The adequacy of a matrix reference material to a sample under analysis (property value match, similarity of matrices and chemical compositions) is one of the basic requirements for the selection and use of reference materials [16]. Direct use of a reference material as a test item for PT without taking account of the adequacy of the material cannot provide traceability and may, particularly in small schemes, lead to totally mistaken results. Therefore, the task of value assignment is divided here into the following three scenarios: (I) an adequate matrix certified reference material (CRM) and/or in-house reference material (IHRM) with traceable property values are available for use as test items; (II) available matrix CRMs are not directly applicable, but a CRM or IHRM can be used in formulating a spiked material with traceable property values and the like; (III) only an IHRM with limited traceability is available (for example, because of instability of the material under analysis).

Scenario I: use of adequate CRM and IHRM

The ideal case is when the test items distributed among the laboratories participating in the PT are portions of an adequate matrix CRM. As a rule, however, CRMs are too expensive for direct use in PT in the capacity of test items, and a corresponding IHRM is to be developed. Characterization of an IHRM with property values traceable to the CRM by comparison, and application of the IHRM for PT are described in [3, 17–21]. The characterization can be effectively carried out by the analysis of the two materials in pairs, each pair consisting of one portion of the IHRM and one portion

of the CRM. A pair is analyzed practically simultaneously, by the same analyst and method, in the same laboratory and conditions. According to this design, the analyte concentration in the IHRM under characterization is compared with the certified value of the CRM and is calculated using differences in the results of the analyte determinations in the pairs. The standard uncertainty of the IHRM certified value is evaluated as a combination of the CRM standard uncertainty and of the differences' standard uncertainty (the standard deviation of the mean of the differences). The uncertainty of the IHRM certified value includes homogeneity uncertainties of both the CRM and the IHRM, since the differences in the results are caused not only by the measurement uncertainties, but also by fluctuations of the analyte concentrations in the test portions. When more than one unit of IHRM is prepared for PT, care still needs to be taken to include the IHRM between-unit homogeneity term in evaluating the uncertainty. Since, in this scenario, the CRM and IHRM have similar matrixes and close chemical compositions, their stability characteristics are assumed to be identical, unless there is information to the contrary. The CRM uncertainty forms part of the IHRM uncertainty budget and is expected to include any necessary uncertainty allowance related to stability, so no additional stability term is included in the IHRM uncertainty [20, 21].

The criterion of the fitness-for-purpose uncertainty of the property value of a reference material applied for PT is formulated depending on the task. For example, for PT in the field of water analysis in Israel, expanded uncertainty values are to be negligible in comparison to the maximum contaminant level (MCL), i.e., the maximum permissible analyte concentration in water delivered to any user of the public water system. In this example, the uncertainty was limited to $2u_{cert}<0.3$ MCL, where 2 is the coverage factor. This limitation can be interpreted in terms of the IUPAC Harmonized Protocol [2] as $u^2_{cert}< 0.1\sigma^2_p$, where σ_p=MCL/2.

Correct planning of the range of analyte concentrations is also important for the scheme. For the example of the water analysis, the suitable range for PT is (0.5–1.5) MCL. The scheme is more effective if two IHRMs with two analyte concentration levels within the range (C_{cert} values lower and higher than MCL) are prepared simultaneously and sent to laboratories as Youden pairs [3, 22].

Scenario II: no closely matched CRMs; IHRM formulation by spiking

The PT scheme for this scenario can be based on a gravimetric preparation of a synthetic IHRM by the addition of a pure substance spike or a traceable but less well matched matrix CRM (further CRM) to a matrix/sample under analysis. For example, a herbicides mixture in acetonitrile is applicable as such a CRM for the preparation of a synthetic water IHRM [21]. The traceable assigned value C_{cert} of the spike in the synthetic IHRM and its standard uncertainty u_{cert} are calculated taking into account: (1) measured masses of the matrix and the CRM; (2) standard uncertainties of the mass measurements and of the analyte concentration certified in the CRM; and (3) standard uncertainty caused by the IHRM homogeneity. The homogeneity can be evaluated by the analysis of the test portions sampled after IHRM preparation (mixing) at the beginning, in the middle, and at the end of the IHRM removal from the mixer into laboratory bottles. The uncertainty component associated with stability is not taken into account if the synthetic IHRM is prepared and used for PT in conditions (temperature, time, etc.), which allow a reasonable assumption that the analyte degradation is negligible.

Approximate preliminary information about the analyte concentration in the matrix/blank (e.g., natural water sample in [21]) and about the analyte total concentration in the synthetic IHRM is necessary only for planning the spike value and afterwards is not so important. In any case, such a blank should have the status of a reference material (with known homogeneity and stability), otherwise, the spike determination will be impossible.

The criterion of fit-for-purpose uncertainty, formulated above for the water analysis, leads here to a similar requirement: the spike expanded uncertainty should be negligible in comparison to the MCL value and should not affect the scoring of the PT results.

A related scenario is based on traceable quantitative elemental analysis and qualitative information on the purity/degradation of the analyte under characterization in the IHRM. For example, IHRMs for the determination of inorganic polysulfides in water have been developed in this way [23]. The determination included the polysulfide's derivatization with a methylation agent followed by gas chromatography/mass spectrometry (GC/MS) or high-pressure liquid chromatography (HPLC) analysis of the difunctionalized polysulfides. Therefore, the IHRMs were synthesized in the form of dimethylated polysulfides containing from four to eight atoms of sulfur. The composition of the compounds was confirmed by nuclear magnetic resonance (NMR) and by the dependence of the HPLC retention time of the dimethylpolysulfides on the number of sulfur atoms in the molecule. Stability of the IHRMs was studied by HPLC with ultraviolet

(UV) detection. The total sulfur content was determined by the IHRM's oxidation with perchloric acid in high-pressure vessels (bombs), followed by the determination of the formed sulfate using inductively coupled plasma–atomic emission spectrometry (ICP-AES). The IHRM certified values are traceable to SI kg, since all of the test portions were weighed, and to the National Institute of Standards and Technology Standard Reference Material (NIST SRM) 682 through the Anion Multi-Element Standard II from ''Merck'' containing sulfate ions of 1,000±5 ppm that was used for the ICP-AES calibration. The chromatographic data were used only for the identification of the polysulfide degradation (as ''yes or no''). The standard uncertainty u_{cert} was about 7.3% rel., which was considered to be sufficient for use in environmental analysis since, given the variation of inorganic polysulfide concentrations in natural water sources, uncertainty up to about 20% rel. is acceptable [23].

Scenario III: appropriate CRMs are not available

This scenario can arise when a component or impurity of an object/material under analysis is unstable, or the matrix is unstable, and no CRMs are available. The proposed PT scheme for such a case is based on the preparation of an individual sample of IHRM for every participant in the same conditions provided by a reference laboratory (RL), allowing the participant to start the measurement/test process immediately after the sample preparation. In this scheme, IHRM instability is not relevant as a source of measurement/test uncertainty, while intra- and between-samples inhomogeneity parameters are evaluated using the results of RL testing of the samples taken at the beginning, the middle, and the end of the PT experiment. For example, such a PT scheme was used for concrete testing [24]. Slump and compressive strength were chosen in the scheme as the test parameters of fresh and hardened concrete, practically the most detail required by the customers.

The concrete for every PT participant (IHRM sample of 35 L) was produced by the RL using the same components, same mixer, and in the same conditions. Every participant had a possibility to start testing its sample from the moment when the concrete preparation was finished. Twenty-five participants took part in the experiment (N=25). Twenty-nine samples were prepared by the RL. The first sample, two in the middle of the experiment, and the last samples were tested by the RL for the material inhomogeneity study and characterization. Other samples were tested by the PT participants according to the schedule prepared and announced in advance.

The slump duplicate determinations were performed at the RL by representatives of every participant using their own facilities and standard operating procedures (SOPs). Immediately after the slump determination, 12 test cubes for compressive strength determinations were prepared by representatives of every participant also using their own facilities and SOPs. On the next day after preparing, the hardened cubes were transferred from the RL to the laboratory of the participant, where compressive strength determinations were performed both on the 7th day and on the 28th day after the sample preparation (every one of six replicates).

The assigned slump and compressive strength values of the IHRM are calculated as averaged RL results. Since the traceability of the assigned values to the international measurement standards and SI units cannot be stated, only local comparability of the results is assessed.

Uncertainties u_{cert} include here the RL measurement/test uncertainty components and the components arising from the material intra-unit and between-units inhomogeneity. Even when inhomogeneity components are statistically insignificant, u_{cert} values could not be negligible, since the RL and participants of the PT used similar measuring instruments and methods, and their measurement/test uncertainties were of the same order. Therefore, u_{cert} values were taken into account for the assessment of performance of participants in the PT scheme using *zeta*-scores [24].

Performance evaluation and scoring

The present IUPAC Harmonized Protocol [2] recommends that z-score values

$$z_i = \frac{x_i - C_{\text{cert}}}{\sigma_{\text{p}}}$$

are considered to be acceptable within ±2, with values outside ±3 unacceptable, and intermediate values questionable (the grounds for that are discussed thoroughly elsewhere [2]). This score provides the simplest and most direct answer to the question: ''is the laboratory performing to the quantitative requirement (σ_{p}) set for the particular scheme?'' The laboratory's quoted uncertainty is not directly relevant to this particular question, so it is not included in the score. Over the longer term, however, a laboratory will score poorly if its real (as opposed to estimated) uncertainty is too large for the job, whether the problem is caused by unacceptable bias or unacceptable variability. This

scoring, based on an externally set value σ_p (without explicitly taking uncertainties of the assigned value and participant uncertainties into account), remains applicable to small schemes, provided that laboratories share a common purpose for which a single value of σ_p can be determined for each round.

Often, however, a small group of laboratories has sufficiently different requirements that a single criterion is not appropriate. It may then (as well as generally) be of interest to consider a somewhat different question about performance: ''are the participant's results consistent with their own quoted uncertainties?'' For this purpose, *zeta* (ζ) and E_n number scores are appropriate. The scores are calculated as:

$$\zeta_i = \frac{x_i - C_{cert}}{\sqrt{u(x_i)^2 + u_{cert}^2}} \text{ and } E_n = \frac{x_i - C_{cert}}{\sqrt{U(x_i)^2 + U_{cert}^2}},$$

where $U(x_i)$ and U_{cert} are expanded uncertainties of the ith participant result x_i and of the certified (or otherwise assigned) value C_{cert}, respectively. *Zeta*-score values are typically interpreted in the same way as z-score values. The E_n number differs from the *zeta*-score in the use of expanded uncertainties and E_n values are usually considered to be acceptable within ±1. The advantages of *zeta*-scoring are that: (1) it takes explicit account of the laboratory's reported uncertainty; (2) it provides feedback on both the laboratory result and on the laboratory's uncertainty estimation procedures. The main disadvantages are that: (1) it cannot be directly related to an independent criterion of fitness-for-purpose; (2) pessimistic uncertainty estimates lead to consistently good *zeta*-scores, irrespective of whether they are fit for a particular task or not; and (3) the PT provider has no way of checking that reported uncertainties are the same as those given to customers, although a customer or accreditation body is able to check this if necessary. The E_n number shares these characteristics, but adds two more. First, it additionally evaluates the laboratory's choice of coverage factor for converting standard to expanded uncertainty—this is an advantage. Second, unless the confidence level is set in advance, E_n is sensitive to the level of confidence chosen both by participant and by provider in calculating $U(x_i)$ and U_{cert}. It is obviously important to ensure consistency in the use of coverage factors if E_n numbers are to be compared.

It is clear that no single score can provide simultaneous information on whether laboratories are meeting external criteria (z-scores apply best here) and on whether they meet their own (*zeta* or E_n number applies best). However, a provider or participant may consider more than one aspect of performance by reviewing two or more scores. It is always possible for a laboratory to calculate its own *zeta*-score if provided with an assigned value and its uncertainty, a point made in detail in [2]. It is less easy for a laboratory to compare with external fitness-for-purpose criteria if they are provided only with a *zeta*-score, but the scheme may additionally set some criteria for either a maximum laboratory uncertainty or for a maximum deviation from the assigned value, which allows a wider fitness-for-purpose judgment.

Effect of a small laboratory population on sample estimates

The population of possible laboratory participants is not usually infinite. For example, the population size of possible PT participants in motor oil testing organized by the Israel Forum of Managers of Oil Laboratories was N_p=12 only, while the sample size, i.e., the number of participants who agreed to take part in the PT in different years was N=6–10 [8]. In such cases, the sample fraction ρ=6/12–10/12=0.5–0.8 (i.e., 50–80%) is not negligible and corrections for finite population size are necessary in some statistical data analyses. The corrections include the standard deviation (standard uncertainty) of the sample mean of N PT results $c_{PT/av}$, equal to $\sigma_{PT/av}=\sigma_{PT}\{[(N_p-N)/(N_p-1)]/N\}^{1/2}$, and the standard deviation of a PT result, equal to $s_{PT}=\sigma_{PT}[N_p/(N_p-1)]^{1/2}$.

After simple transformations, the following formula for the sample mean can be obtained: $\sigma_{PT/av}/(\sigma_{PT}/\sqrt{N})=[(N_p-N)/(N_p-1)]^{1/2}=[(1-\rho)/(1-1/N_p)]^{1/2}$. The dependence of $\sigma_{PT/av}$ on ρ, %, is shown (in units of $\sigma_{PT}/\sqrt{N}$) in Fig. 3 for the populations of N_p=10, 20, and 100 laboratories; curves 1, 2, and 3, respectively. Since at least two PT results are necessary for the calculation of a standard deviation (i.e., the minimal sample size is N=2), curve 1 is shown for $\rho\geq$20%, curve 2 for $\rho\geq$10%, and curve 3 for $\rho\geq$2%. The population size has much less influence here than the sample fraction value. The dependence of s_{PT} on ρ, %, by the formula $s_{PT}/\sigma_{PT}=[1/(1-\rho/N)]^{1/2}$ is shown (in σ_{PT} units) in Fig. 4 for the sample sizes of N=10, 20, and 100 PT results; curves 1, 2, and 3, respectively. This dependence is not as dramatic as the previous one, since the correction factor values (the ordinate range) are of 0.96–1.00 only for any event. As N_p increases and ρ decreases, the values $(N_p-N)/(N_p-1)\rightarrow 1$ and $1/(1-\rho/N)\rightarrow 1$, and the corrections for finite population size disappear: $\sigma_{PT/av}\rightarrow\sigma_{PT}/\sqrt{N}$ and $s_{PT}\rightarrow\sigma_{PT}$ [7]. Therefore, the

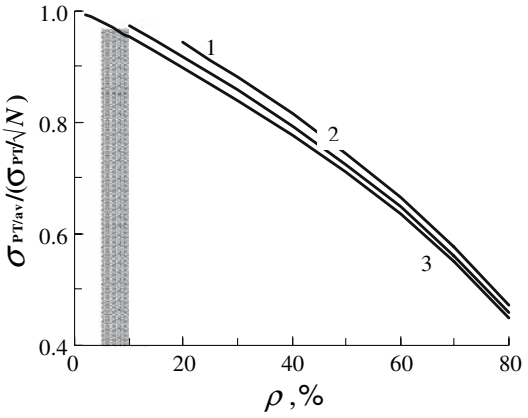

Fig. 3 Dependence of the standard deviation of the sample mean $\sigma_{PT/av}$ on the sample fraction ρ, %, in units of $\sigma_{PT}/\sqrt{N}$. Curves 1, 2, and 3 are for the populations of N_p=10, 20, and 100 laboratories, respectively. The *gray bar* shows the intermediate range of sample fraction values ρ=5–10% (at ρ<5–10%, corrections for a finite population size are negligible, as a rule)

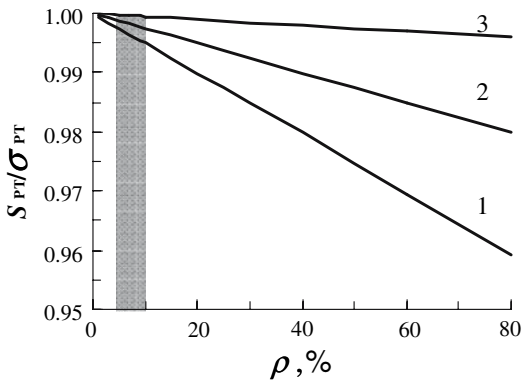

Fig. 4 Dependence of the standard deviation of a PT result s_{PT} on the sample fraction ρ, %, in units of σ_{PT}. Curves 1, 2, and 3 are for the sample sizes of N=10, 20, and 100 PT results, respectively. The intermediate range of sample fraction values ρ=5–10% is shown by the *gray bar*

corrections are negligible for ρ values up to around 5–10% (shown by the gray bars in Figs. 3 and 4).

These corrections should, however, be applied with care, and only when the population is really finite. They do not apply, for example, if the errors for laboratories are predominantly random from round to round; this would imply that, while the population of laboratories is finite, each round effectively provides a sample of laboratory errors from an infinite population of errors.

Since the number of PT results (the sample size N) is limited, it is also important to treat extreme results correctly if they are not caused by a known gross error or miscalculation. Even at large N, extreme results can provide valuable information to the PT provider and should not be disregarded entirely in the analysis of the PT results without due consideration [25]. If N is small, extreme results cannot usually be identified as outliers by known statistical criteria because of their low power [7]. Fortunately, the metrological approach for small schemes makes outlier handling less important, since assigned values should not be calculated by consensus, and scores are not expected to be based on observed standard deviations. Outliers, accordingly, have an effect on scoring only for the laboratory reporting outlying results and for the PT provider seeking the underlying causes of such problems. Scores should, in any case, be provided to all participants, whether or not their individual results are extreme.

Conclusions

1. Consensus mean values and observed standard deviations of measurement/analytical results of laboratories participating in proficiency testing (PT) are insufficiently reliable for the assessment of a laboratory's performance in a PT with a limited number of participants (less than 20–30).
2. Traceable assigned values of test items (portions of a certified reference material, of an in-house reference material or of a spike) and externally set performance criteria (usually expressed as a standard deviation of PT results for a z-score) acceptable for all participants should be used wherever at all possible. When information necessary to set external performance criteria is not available, the assigned value uncertainty is not negligible, or the laboratories are working according to their own fitness-for-purpose criteria (a single criterion is inapplicable for all participants), the information included in the measurement uncertainties reported by the laboratories may be helpful for their proficiency assessment with *zeta*-score the or E_n number. In general, a wider range of scoring methods is likely to be appropriate to meet the needs of all participants. An optimal PT scheme is to be selected depending on existing reference materials or on the ability to develop such materials for the PT purposes and on suitable scores for the participant performance assessment.
3. Statistical methods applied for treatment of the PT results (including the detection of outliers) should take into account a small size of the sample derived from the infinite population of possible PT participants. If the population size is also limited, a correction for the sample fraction may be necessary.

References

1. Thompson M, Wood R (1993) Pure Appl Chem 65(9):2123–2144
2. Thompson M, Ellison SLR, Wood R (2006) Pure Appl Chem 78(1):145–196
3. ISO 13528:2005 (2005) Statistical methods for use in proficiency testing by interlaboratory comparisons
4. 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
5. ILAC G13 (2000) Guidelines for the requirements for the competence of providers of proficiency testing schemes
6. Deom A, Aouad REl, Heuck CC, Kumari S, Lewis SM, Uldall A, Wardle AJ (1999) Requirements and guidance for external quality assessment schemes for health laboratories. World Health Organization, WHO/DIL/LAB/99.2
7. Dixon WJ, Massey FJ Jr (1969) Introduction to statistical analysis, 3rd edn. McGraw-Hill, New York, pp 111–113
8. Kardash-Strochkova E, Tur'yan YaI, Kuselman I, Brodsky N (2002) Accred Qual Assur 7:250–254
9. IUPAC project 2005-019-2-500. Selection and use of proficiency testing schemes for limited number of participants (chemical analytical laboratories). Description online at http://www.iupac.org/projects/2005/2005-019-2-500.html
10. EURACHEM/CITAC guide (2003) Traceability in chemical measurement. A guide to achieving comparable results in chemical measurement
11. ISO/IEC 17025:2005 (2005) General requirements for the competence of testing and calibration laboratories
12. Armishaw P, King B, Millar RG (2003) Accred Qual Assur 8:184–190
13. Kuselman I (2006) Accred Qual Assur 10:466–470
14. Kuselman I (2006) Accred Qual Assur 10:659–663
15. Kuselman I (2007) In: Fajgelj A, Belli M, Sansone U (eds) Combining and reporting analytical data. RCS special publication No. 307, Cambridge, UK
16. Kuselman I (2004) Accred Qual Accur 9:591–596
17. Kuselman I, Weisman A, Wegscheider W (2002) Accred Qual Assur 7:122–124
18. Ekeltchik I, Kardash-Strochkova E, Kuselman I (2003) Microchim Acta 141:195–199
19. Weisman A., Gafni Y, Vernik M, Kuselman I (2003) Accred Qual Assur 8:263–266
20. Kuselman I (2004) J Metrol Soc India 19(4):245–252
21. Kuselman I, Pavlichenko M (2004) Accred Qual Accur 9:387–390
22. Youden WJ, Steiner EH (1990) Statistical manual of the Association of Official Analytical Chemists. AOAC International, Arlington, Virginia
23. Rizkov D, Lev O, Gun J, Anisimov B, Kuselman I (2004) Accred Qual Assur 9:399–403
24. Kimhi L, Zlotnikov C, Kuselman I (2006) Accred Qual Assur 11:577–583
25. Kuselman I (1999) Accred Qual Assur 4:511