Pure Appl. Chem., ASAP Article
doi:10.1351/PAC-REP-09-08-15

IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for a limited number of participants—chemical analytical laboratories (IUPAC Technical Report)*

Ilya Kuselman[1,‡] and Aleš Fajgelj[2]

[1]The National Physical Laboratory of Israel, Givat Ram, Jerusalem 91904, Israel; [2]International Atomic Energy Agency, Wagramer Strasse 5, P.O. Box 100, Vienna A-1400, Austria

Abstract: A metrological background for implementation of proficiency testing (PT) schemes for a limited number of participating laboratories (fewer than 30) is discussed. Such schemes should be based on the use of certified reference materials (CRMs) with traceable property values to serve as PT items whose composition is unknown to the participants. It is shown that achieving quality of PT results in the framework of the concept "tested once, accepted everywhere" requires both metrological comparability and compatibility of these results.

The possibility of assessing collective/group performance of PT participants by comparison of the PT consensus value (mean or median of the PT results) with the certified value of the test items is analyzed. Tabulated criteria for this assessment are proposed.

Practical examples are described for illustration of the issues discussed.

Keywords: IUPAC Analytical Chemistry Division; measurement uncertainty; metrological comparability; metrological compatibility; metrological traceability; proficiency testing; sample size.

CONTENTS

*Sponsoring bodies: IUPAC Analytical Chemistry Division; IUPAC Interdivisional Working Party on Harmonization of Quality Assurance; Cooperation on International Traceability in Analytical Chemistry (CITAC): see more details on p. xxxx.

‡Corresponding author: E-mail: ilya.kuselman@moital.gov.il

1. INTRODUCTION

The International Harmonized Protocol for proficiency testing (PT) of analytical chemistry laboratories adopted by IUPAC in 1993 [1] was 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]. General requirements for PT are updated in the new standard [5]. *International Laboratory Accreditation Cooperation (ILAC) Guidelines* define requirements for the competence of PT providers [6]. Guidelines for PT use in specific sectors, like clinical laboratories, have also been widely available [7]. In some other sectors, they are under development.

These documents are, however, oriented mostly toward PT schemes for a relatively large number N of laboratories or participants (greater than or equal to 30), henceforth referred to as "large schemes". This is important from a statistical point of view, since with $N < 30$, evaluations by statistical methods become increasingly unreliable, especially for $N < 20$. For example, uncertainties in estimates of location (such as mean and median) are sufficiently small to be neglected in scoring as N increases to approximately 30, but cannot be neglected safely with $N < 20$. Deviations from a normal distribution are harder to identify if N is small. Robust statistics are not usually recommended when $N < 20$. Therefore, the assigned/certified value of the PT items c_{cert} cannot be calculated safely from the measurement results obtained by the participants (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 (fewer than 30). The intermediate "gray" range of $20 \le N < 30$ is included in the range of small schemes since such N values may influence PT planning and interpretation also.

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 greater than 5–10 % of N_{p}, the value of the sample fraction $\varphi = N/N_{\mathrm{p}}$ may need to be taken into account.

Thus, implementation of small PT schemes is sometimes not a routine task. Such schemes are quite often required for quality assurance of environmental analysis specific for a local region, analysis of specific materials in an industry (e.g., under development), for purposes of a regulator or a laboratory accreditation body, etc. [8].

1.1 Scope and field of application

This Guide is developed for implementation of simultaneous participation schemes when the number of laboratories is smaller than 30. This includes: (1) selection of a scheme based on simultaneous distribution of test items to participants for concurrent quantitative testing; (2) use of certified reference

materials (CRMs) as test items unknown to the participants; (3) the individual laboratory performance assessment and assessment of the metrological comparability and compatibility of the measurement results of the laboratories taking part in the PT scheme as a collective (group) of the participants.

The document is intended for PT providers and participants (chemical analytical laboratories), for accreditation bodies, laboratory customers, regulators, quality managers, metrologists, and analysts.

1.2 Terminology

Terminology used in this Guide corresponds to ISO standards 17043 [5] and 3534 [9], and ISO Guide 99 (VIM3) [10].

2. APPROACH

2.1 Properties of PT consensus values: Dependence on statistical sample size

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 0.95 probability, 1.96 being the appropriate quantile of the normal distribution for a two-sided 0.95 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 indicated by the gray bar. Even for N = 30, the bias may reach 0.36 σ_{PT} at the 0.95 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 probability of 0.95, where $\chi^2\{\alpha, N-1\}$ is the α quantile 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 0.95 limit for s_{PT} is 1.26σ_{PT}. In other words, s_{PT} can differ from σ_{PT} for N = 30 by over 25 % rel. at the level of confidence 0.95. For N < 30, the difference between the sample and the population characteristics increases with decreasing N, especially dramatically for the standard deviation when N < 20.

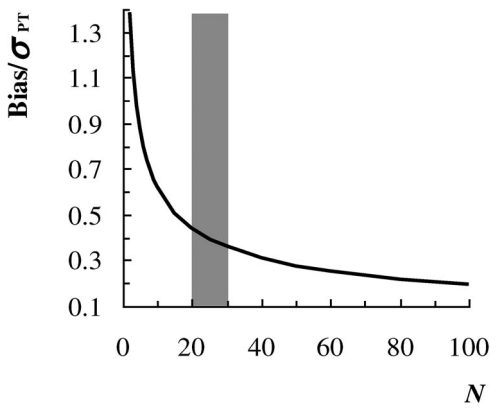

Fig. 1 Dependence of the upper limit of the bias $|c_{PT/avg} - c_{PT}|$ (in units of σ_{PT}) on the number N of PT results; reproduced from ref. [8] by permission of Springer. The line is the 0.975 quantile corresponding to the upper limit of the two-sided 0.95 interval for the expected bias. The intermediate range of N = 20 to 30 is shown by the gray bar.

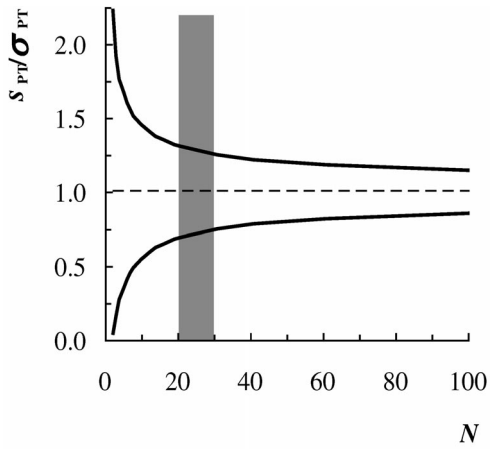

Fig. 2 Dependence of the sample standard deviation s_{PT} limits (in units of σ_{PT}) on the number N of PT results; reproduced from ref. [8] by permission of Springer. Solid lines show 0.025 (lower line) and 0.975 (upper line) quantiles for s_{PT}. The dashed line is at $s_{PT}/\sigma_{PT} = 1.0$ for reference. The gray bar shows the intermediate range of sample sizes $N = 20$ to 30.

While consensus mean values are less affected than observed standard deviations, uncertainties in 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] for cases when the uncertainties are negligible. It follows that scoring for small schemes should usually avoid simple consensus values. Methods of obtaining traceable assigned values c_{cert} are to be used wherever possible to provide comparable PT results [11,12].

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 not recommended even for large schemes [3], on the grounds that it does not provide 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 an external, fit-for-intended-use, normative or target standard deviation σ_{targ}, then z-scores, which compare a result bias from the assigned value with σ_{targ}, can be calculated in a small scheme in the same manner as recommended in refs. [1–5] for a large scheme. The condition is only that the standard uncertainty of the assigned/certified value u_{cert} is insignificant in comparison to σ_{targ} ($u_{cert}^2 < 0.1\sigma_{targ}^2$).

2.2 Measurement uncertainty use for interpretation of PT results

When information necessary to set σ_{targ} is not available, and/or u_{cert} is not negligible, the information, included in the measurement uncertainty $u(c_i)$ of the result c_i reported by the i-th laboratory, is helpful for performance assessment using ζ-scores and/or E_n numbers [2,3]. It may also be important for a small scheme that laboratories working according to their own fit-for-intended-use criteria (e.g., in conditions of competition) can be judged by individual criteria based on their declared measurement uncertainty values.

2.3 What is a metrological approach to PT?

The approach based on metrological traceability of an assigned value of test items, providing comparability of PT results, and on scoring PT results taking into account uncertainties of the assigned value and uncertainties of the measurement results, has been described as a "metrological approach" [13].

Two main steps are common for any PT scheme using this approach: (1) establishment of a metrologically traceable assigned value, c_{cert}, of analyte concentration in the test items/RM and quantification of the standard uncertainty u_{cert} of this value, including components arising from the material homogeneity and stability during the PT round, and (2) calculation of fit-for-intended-use performance statistics as well as assessment of the laboratory performance, taking into account the laboratory measurement uncertainty. For the second step, it may be necessary in addition to take into account the small population size of laboratories able to take part in the PT. These issues are considered below.

3. VALUE ASSIGNMENT

3.1 Metrological traceability of a CRM property value and of PT results

Since the approach to PT for a limited number N of participants is based on the use of CRMs as test items unknown to the participants, metrological traceability of a CRM property value is a key to understanding metrological comparability and compatibility of the PT results. Interrelations of these parameters are shown in Fig. 3. The left pyramid in Fig. 3 illustrates the calibration hierarchy of CRMs as measurement standards or calibrators [10] ranked by increasing uncertainties of supplied property values from primary CRMs (mostly pure substances developed by National Metrology Institutes, NMIs), to secondary CRMs (e.g., a matrix CRM traceable to primary CRMs), and from secondary to

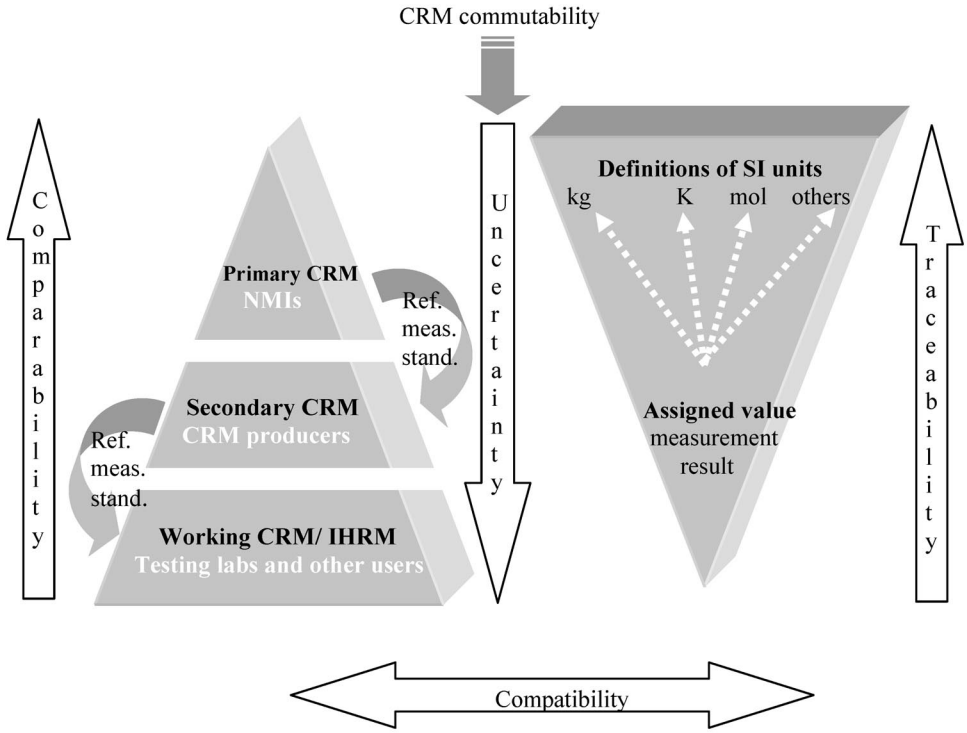

Fig. 3 A scheme of calibration hierarchy, traceability and commutability (adequacy or match) of RMs used for PT, comparability and compatibility of PT results; reproduced from ref. [16] by permission of Springer.

working CRMs (certified in-house RMs, IHRMs, developed by testing/analytical laboratories, PT providers, and other users) [14,15]. When a CRM of a higher level is used for certification of an RM of a lower level by comparing them (e.g., for certification of IHRM), the first one plays the role of a reference measurement standard, shown in Fig. 3 by semicircular pointers. Since uncertainty of CRM property values is increasing in this way, the uncertainty pointer is directed from the top of the pyramid to the bottom.

The same CRM can be used for calibration of a measurement system and for PT, i.e., for two different purposes: as a calibrator and as a quality control material (test items), but not at the same time, in the same measurement or in the same test [17].

The right-side overturned pyramid in Fig. 3 shows traceability chains from an RM certified value and the corresponding measurement/analysis/test results to definitions of SI units. As a rule, one result is to be traceable to the definition of its unit, while simultaneously there are several influence quantities which need also to be traceable to their own definition of units: to the mole of the analyte entities per mass of sample (i.e., for the concentrations in the calibration solutions), to the kilogram or meter because the size of a sample under analysis is quantified by mass or volume, to the kelvin when the temperature influences the results obtaining for the main quantity, etc. Thus, the traceability pointer has a direction which is opposite to the measurement uncertainty. Of course, the width of the overturned pyramid is not correlated with the uncertainty values, as is the case in the left-side pyramid.

Understanding traceability of measurement/analysis/test and PT results to the mole (realized through the chain of the CRMs according to their hierarchy) is often not simple and requires reliable information about the measurement uncertainty. The problem is that the uncertainty of analytical results may increase because of deviations of the chemical composition of the matrix CRM (used for calibration of the measurement system) from the chemical composition of the routine samples under analysis. Similarly, the difference between a certified value of the matrix RM (applied in a PT as test items) and the result of a laboratory participating in the PT may increase when the CRM has a different chemical composition than the routine samples. This is known as the problem of CRM commutability, adequacy or match, to a sample under analysis [18], and is shown in Fig. 3 as an additional pointer above the uncertainty pointer. The commutability is discussed in the following paragraph 3.1.1, while the metrological comparability and compatibility pointers, shown also in Fig. 3, are described in paragraph 5.

3.1.1 Commutability of the CRMs and routine samples

Since a difference in property values and matrices of CRM and of routine samples influences the measurement uncertainty in PT, the chemical composition of both, the measurement standard (the CRM used as test items) and the routine samples of the test object, should be as close as possible. An algorithm for *a priori* evaluation of CRM adequacy can be based on the use of an adequacy score: $A_S = \prod_i^n R_i^{a_i}$, where Π is the product, $i = 1, 2, \ldots, n$ is the number of a component or of a physico-chemical parameter; $R_i = [\min(c_{i,s}, c_{i,cert})/\max(c_{i,s}, c_{i,cert})]$ is the ratio of the minimum to the maximum values from $c_{i,s}$ and $c_{i,cert}$; $c_{i,s}$ and $c_{i,cert}$ are the concentrations of the i-th component or the values of the i-th physico-chemical parameter in the sample and certified in the CRM, respectively; $0 \le a_i \le 1$ is the empirical sensitivity coefficient which allows decreasing the influence of a component or a parameter on the score value, if the component or the parameter is less important for the analysis than others. According to this score, the ideal adequacy ($A_S = 1$ or 100 %) is achieved when the composition and properties of the sample and of the RM coincide. The adequacy is absent ($A_S = 0$ %) when the sample and the CRM are different substances or materials, and/or the analyte is absent in the CRM ($c_{i,cert} = 0$). Intermediate cases, for example, for two components under control are shown in Fig. 4. The ratios R_1 and R_2 providing adequacy score values $A_S = 70$, 80, and 90 %, form here curves 1–3, respectively.

The adequacy score may be helpful for CRM choice as a calibrator since direct use of a CRM having a low adequacy score can lead to an incorrect/broken traceability chain. Such a CRM applied for PT will decrease the reliability of a laboratory performance assessment. Therefore, CRM commutability in PT and a score allowing its evaluation are also important. However, the adequacy score

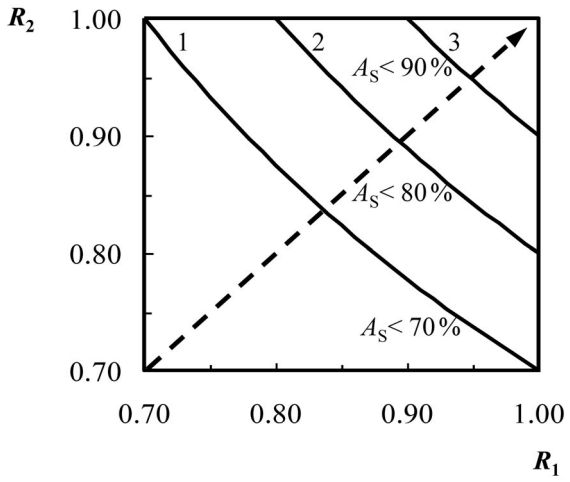

Fig. 4 Adequacy score A_S values in dependence on ratios R_1 and R_2 of concentrations of two components in a sample under analysis and in a CRM; reproduced from ref. [16] by permission of Springer. Curves 1–3 correspond to A_S = 70, 80 and 90 %, respectively. The dotted pointer shows the direction of the adequacy increasing.

does not properly quantify the measurement uncertainty contribution caused by insufficient commutability (A_S < 100 %). This requires a special study.

For more details of A_S calculations, see Example 5 in Annex B.

3.1.2 Three scenarios

Thus, the task of value assignment is divided into the following three scenarios: (I) an adequate matrix CRM with traceable property value is available for use as test items; (II) available matrix CRMs are not directly applicable, but a CRM can be used in formulating a spiked material with traceable property values; (III) only an IHRM with a limited traceability chain of the property value is available (e.g., owing to instability of the material under analysis).

3.2 Scenario I: Use of adequate CRM

The ideal case is when the test items distributed among the laboratories participating in the PT are portions of a purchased adequate matrix CRM (primary or secondary measurement standard). However, when the CRMs available in the market are too expensive for direct use in PT in the capacity of test items, a corresponding IHRM (working measurement standard) is to be developed. Characterization of an IHRM with a property value traceable to the CRM value by comparison, and application of the IHRM for PT are described in refs. [3,19–21]. The characterization can be effectively carried out by 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 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 standard uncertainty of the differences (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 measured 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 uncer-

tainty. Since, in this scenario, the CRM and IHRM have similar matrixes and close chemical compositions, at similar processing, packaging, and transportation conditions their stability characteristics during PT are assumed to be identical unless there is information to the contrary. The CRM uncertainty forms a part of the IHRM uncertainty budget and is expected to include any necessary uncertainty related to stability, therefore, no additional stability term is included in the IHRM uncertainty.

The criterion of fit-for-intended-use uncertainty of the property value of an RM applied for PT is formulated depending on the task. For example, for PT in the field of water analysis in Israel [22], expanded uncertainty values should 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 $u_{cert}{}^2 < 0.1\ \sigma_{targ}{}^2$, where σ_{targ} = 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 [3,23].

For a more detailed example, see Example 1 in Annex B.

3.3 Scenario II: No closely matched CRMs

The PT scheme for this scenario can be based on a gravimetric preparation of a synthetic IHRM by addition of a spike (pure substance or a less well matched matrix CRM available in the market) to a matrix/sample under analysis. For example, a mixture of herbicides in acetonitrile is applicable as such a CRM for preparation of IHRM, synthetic water [22].

Approximate preliminary information about the analyte concentration in the matrix/blank (e.g., natural water sample) and about the analyte total concentration in the synthetic IHRM is necessary only for planning the spike value. In any case, such a blank should have the status of an RM with known homogeneity and stability, otherwise the spike determination will be impossible.

The criterion of fit-for-intended-use uncertainty, formulated above for 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 purity/degradation of the analyte under characterization in the IHRM. For example, IHRMs for determination of inorganic polysulfides in water have been developed in this way [24]. The determination included the derivatization of polysulfides with a methylation agent followed by gas chromatography-mass spectroscopy (GC-MS) or high-performance liquid chromatography (HPLC) analysis of the difunctionalized polysulfides. Therefore, the IHRMs were synthesized in the form of dimethylated polysulfides containing four to eight atoms of sulfur. Composition of the compounds was confirmed by nuclear magnetic resonance (NMR) and by dependence of 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. Total sulfur content was determined by the IHRMs' oxidation with perchloric acid in high-pressure vessels (bombs), followed by determination of the sulfate formed using inductively coupled plasma-optical emission spectroscopy (ICP-OES). IHRM certified values were traceable to NIST SRM 682 through the Anion Multi-Element Standard II from "Merck" (containing certified concentration of sulfate ions) that was used for the ICP-OES calibration, and to the SI kg, since all the test portions were quantified by weight.

For a more detailed example, see Example 2 in Annex B.

3.4 Scenario III: Appropriate CRMs are not available

This scenario can arise when a component or an impurity of an object/material under analysis is unstable, or the matrix is unstable, and no CRMs (primary or secondary measurement standards) are available. The proposed PT scheme for such a case is based on 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: more details, see Example 3 in Annex B.

4. INDIVIDUAL LABORATORY PERFORMANCE EVALUATION AND SCORING

4.1 Single (external) criterion for all laboratories participating in a PT

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

$$z_i = \frac{c_i - c_{\text{cert}}}{\sigma_{\text{targ}}}$$

are considered acceptable within ±2, unacceptable with values outside ±3, and questionable with intermediate values. The grounds for these criteria 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_{targ}) 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 be scored 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_{targ} (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_{targ} can be determined for each round.

For examples of the σ_{targ} setting and z-score use, see Examples 1–2 in Annex B.

4.2 Own criterion for every laboratory

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

$$\zeta_i = \frac{c_i - c_{\text{cert}}}{\sqrt{u(c_i)^2 + u_{\text{cert}}^2}} \quad \text{and} \quad E_n = \frac{c_i - c_{\text{cert}}}{\sqrt{U(c_i)^2 + U_{\text{cert}}^2}}$$

where $u(c_i)$ and $U(c_i)$ are the standard and expanded uncertainties of the i-th participant result c_i, respectively, U_{cert} is the expanded uncertainty of the certified (or otherwise assigned) value c_{cert}. ζ-score values are typically interpreted in the same way as z-score values (see Annex B, Example 3). E_n number differs from ζ-score in the use of expanded uncertainties and E_n values are usually considered acceptable within ±1. The advantages of ζ-scoring are that (i) it takes explicit account of the laboratory's reported uncertainty; (ii) it provides feedback on both the laboratory result and on the laboratory's uncertainty estimation procedures. The main disadvantages are that (i) it cannot be directly related to an independent criterion of fitness-for-intended-use; (ii) pessimistic uncertainty estimates lead to consis-

tently good ζ-scores irrespective of whether they are fit for a particular task; and (iii) 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. Secondly, unless the confidence level is set in advance, E_n is sensitive to the confidence level chosen both by participant and by provider in calculating $U(c_i)$ 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 a single score cannot provide simultaneous information on whether laboratories meet external criteria (where z-scores apply best) and on whether they meet their own criteria (where ζ or E_n number apply best).

5. METROLOGICAL COMPARABILITY AND COMPATIBILITY OF PT RESULTS

The meaning of metrological comparability of PT results is that, being traceable to the same metrological reference, they are comparable independently of the result values and of the associated measurement uncertainties. Since scoring a laboratory proficiency in the discussed small PT schemes is based on evaluation of the bias $|c_i - c_{\text{cert}}|$ of i-th laboratory result c_i from the certified property value c_{cert} of the test items, both PT results and the CRM certification (measurement) data should be comparable, i.e., traceable to the same metrological reference. The same is correct for different runs of the PT scheme, when laboratory score values obtained in these runs are compared. As much as metrological comparability is a consequence of metrological traceability, the comparability pointer in Fig. 3 is directed in the same way as the traceability pointer.

Metrological compatibility can be interpreted for PT results as the property satisfied by each pair of PT results, so that the absolute value of the difference between them is smaller than some chosen multiple of the standard measurement uncertainty of that difference. Moreover, successful PT scoring means that the absolute value of the bias $|c_i - c_{\text{cert}}|$ is smaller than the corresponding chosen multiple of the bias standard uncertainty. In other words, a PT result is successful when it is compatible with the CRM (test item) certified value. Therefore, compatibility is shown in Fig. 3 by a horizontal pointer uniting the direct and the inversed pyramids.

Thus, achieving the quality of measurement/analysis/test and PT results in the framework of the concept "tested once, accepted everywhere" [11,25] requires both comparability and compatibility of the results.

When PT is based on the metrological approach, there are two key parameters for assessment of comparability and compatibility of results [26]: (1) position of the CRM sent to the participants in the calibration hierarchy of measurement standards, and (2) closeness of the distribution of PT results to the distribution of the CRM data.

The position of a CRM in the calibration hierarchy depends on the top measurement standard in the traceability chain. For example, if a CRM property value is traceable to definitions of SI units (by scenarios I and II), it confirms world-wide comparability of PT results. Any PT scheme based on the use of IHRM with a limited traceability chain of the property value (not traceable to definitions of SI units: scenario III) provides the possibility of confirming local comparability only. The same situation occurred in the classical fields of mass and length measurements before the Convention of the Metre, when measurement results in different countries had been traceable to different national (local) measurement standards.

At any traceability of the CRM property value used, the closeness of the distributions of the PT results and of the CRM data is important for the result compatibility and performance assessment. Since laboratory performance is assessed individually for each PT participant, even in a case when the performance of the majority of them is found to be successful, compatibility of all the PT results (i.e., a group performance characteristic of the laboratories participating in PT) still remains unassessed.

The situation is illustrated in Fig. 5, where both distribution density functions f of PT results (curve 1) and of CRM data (curve 2) are shown as normal ones. The vertical lines are the centers of these distributions: c_{PT} and c_{cert}, respectively. The common shaded area P under the density function curves is the probability of obtained PT results belonging to the population of the RM data. It can be considered as a parameter of compatibility. The value P tends to zero when the difference between c_{PT} and c_{cert} is significantly larger than standard deviations σ_{PT} and u_{cert} of both distributions. The closer c_{PT} is to c_{cert} (shown by the semicircular pointers in Fig. 5), the higher the P value is.

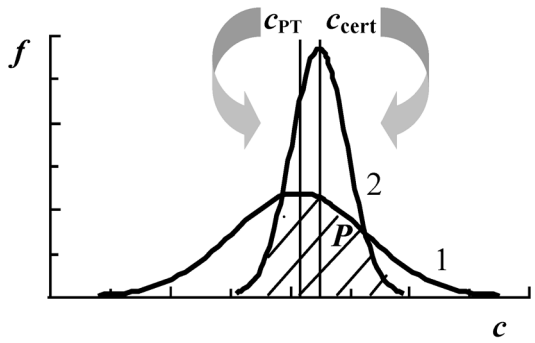

Fig. 5 Probability density functions f of PT results, curve 1, and of CRM data, curve 2; reproduced from ref [16] by permission of Springer. Vertical lines are the centers of these distributions: c_{PT} and c_{cert}, respectively. The common shaded area under the density function curves is the probability P of obtained PT results belonging to the population of the CRM data. The semicircular pointers show the direction of the compatibility increasing.

The distributions, P values, hypotheses necessary for assessment of compatibility of results of a limited number N of PT participants, as a group, and suitable criteria for that based on analysis of the statistical sample characteristics (average $c_{PT/avg}$, standard deviation s_{PT}, etc.) are discussed in detail in Annex A.

In principle, $c_{PT/avg}$ and s_{PT} are the consensus values that cannot be used for a reliable assessment of an individual laboratory performance when the number of the laboratories participating in the PT scheme is limited. However, here the consensus values are used for another purpose: for comparison of PT results, as a statistical sample, with the CRM data (see Examples 1–4 in Annex B). The compatibility of PT results of a group of laboratories can be low if one or more laboratories from the group perform badly. Analysis of reasons leading to such a situation, as well as ways to correct it, are a task for the corresponding accreditation body and/or the regulator responsible for these laboratories and interested in the comparability and compatibility of the results.

6. EFFECT OF 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 used motor oil testing organized by the Israel Forum of Managers of Oil Laboratories was $N_p = 12$ only, while the statistical sample size, i.e., the number of the participants agreed to take part in the PT in different years was $N = 6$–10 (see Annex B, Example 4). In such cases, the sample fraction $\varphi = 6/12$ to $10/12 = 0.5$ to 0.8 (i.e., 50–80 %) is not negligible and corrections for finite population size are necessary in the 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 standard deviation of the sample mean can be obtained: $\sigma_{PT/av}/(\sigma_{PT}/\sqrt{N}) = [(N_P - N)/(N_P - 1)]^{1/2} = [(1 - \varphi)/(1 - 1/N_p)]^{1/2}$. The dependence of $\sigma_{PT/av}$ on φ is shown (in units of $\sigma_{PT}/\sqrt{N}$) in Fig. 6 for the populations of N_P = 10, 20, and 100 laboratories, curves 1–3, respectively.

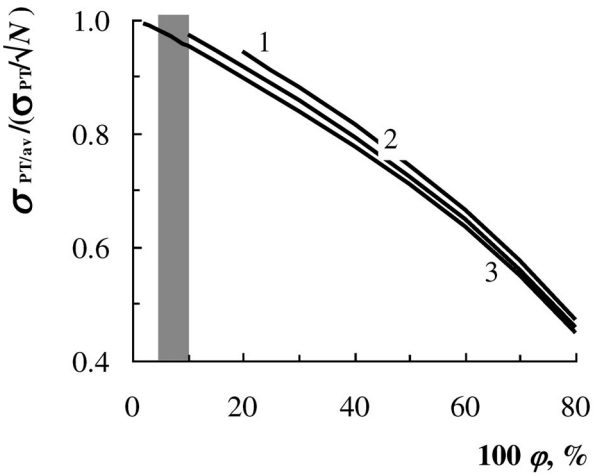

Fig. 6 Dependence of the standard deviation of the sample mean $\sigma_{PT/av}$ (in units of $\sigma_{PT}/\sqrt{N}$) on the sample fraction φ, reproduced from ref. [8] by permission of Springer. Curves 1, 2 and 3 are for the populations of N_P = 10, 20, and 100 laboratories, respectively. The gray bar shows the intermediate range of sample fraction values φ = 5 to 10 % (at φ < 5 % corrections for a finite population size are negligible, as a rule).

Since at least two PT results are necessary for calculation of a standard deviation (i.e., the minimal sample size is N = 2), curve 1 is shown for $\varphi \geq 20$ %, curve 2 for $\varphi \geq 10$ %, and curve 3 for $\varphi \geq 2$ %. The population size has much less influence here than the sample fraction value.

Dependence of s_{PT} on φ by the formula $s_{PT}/\sigma_{PT} = [1/(1 - \varphi/N)]^{1/2}$ is weak in comparison with the previous one in Fig. 6, since the correction factor values are of 0.96–1.00 only for any event when the sample size is of N = 10–100 PT results.

As N_P increases and φ decreases, the values $(N_P - N)/(N_P - 1) \rightarrow 1$ and $1/(1 - \varphi/N) \rightarrow 1$, and the corrections for finite population size disappear: $\sigma_{PT/av} \rightarrow \sigma_{PT}/\sqrt{N}$ and $s_{PT} \rightarrow \sigma_{PT}$. Therefore, the corrections are negligible for φ values up to around 5–10 % (shown by the gray bars in Fig. 6).

These corrections should, however, be applied with care, only when the population is really finite.

7. OUTLIERS

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 analysis of the PT results without due consideration. When N is small, extreme results cannot usually be identified as outliers by known statistical tests because of the low power of these tests.

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. Accordingly, outliers have effect on scoring only for the laboratory reporting outlying results and for the PT provider seeking the underlying causes of such problems.

8. EFFECTIVENESS OF APPROACHES TO PT

While traditional approaches to PT (used consensus values for assessment of a laboratory performance) are not acceptable for $N < 30$, the metrological approach (based on the CRM use) is acceptable from statistical and metrological points of view for any N, including $N \geq 30$ as well. However, a PT cost increasing with N should also be taken into account for any correct PT scheme design.

ANNEX A. CRITERIA FOR ASSESSMENT OF METROLOGICAL COMPATIBILITY OF PT RESULTS

CONTENTS

A-1. RELATIONSHIP BETWEEN THE DISTRIBUTION OF CRM ASSIGNED VALUE DATA AND THE DISTRIBUTION OF PT RESULTS

Data used for calculation of the CRM assigned value and the measurement/analysis results of the laboratories participating in PT can be considered as independent random events. Therefore, the relation between them can be characterized by the common area P under the density function curves for both CRM data and for PT results. The P value is the probability of joint events and, therefore, the probability of PT results obtained belonging to the population of CRM data.

For the sake of simplicity, both distributions are assumed to be normal, with parameters c_{cert}, σ_{cert} and c_{PT}, σ_{PT}, as shown in Fig. 7. The figure refers to a simulated example of aluminum determination in coal fly ashes using a CRM developed by NIST, USA: SRM 2690 with $c_{cert} = 12.35$ % and $\sigma_{cert} = 0.14$ % as mass fraction [27].

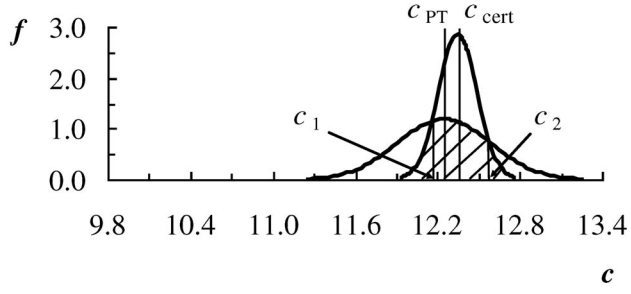

Fig. 7 Probability density functions f of the PT results and of the CRM data when $c_{PT} = 12.25$ % and $\sigma_{PT} = 0.34$ %; reproduced from ref. [27] by permission of RSC. Values c_1 and c_2 are the measurement/test results corresponding to the crossing points of the f curves.

Since both density functions, f_{cert} of CRM data and f_{PT} of PT results are equal at the c_1 and c_2 values, one can write

$$f_{PT} = \frac{1}{\sigma_{PT}\sqrt{2\pi}} \exp\left[-(c - c_{PT})^2 / 2\sigma_{PT}^2\right] = \frac{1}{\sigma_{cert}\sqrt{2\pi}} \exp\left[-(c - c_{cert})^2 / 2\sigma_{cert}^2\right] = f_{cert} \quad (1)$$

As shown in ref. [27], after transformations of expression 1, c_1 and c_2 can be calculated by the following formula:

$$c_1, c_2 = \frac{\left(\sigma_{cert}^2 c_{PT} - \sigma_{PT}^2 c_{cert}\right) \pm \sigma_{cert}\sigma_{PT}\sqrt{\delta}}{\sigma_{cert}^2 - \sigma_{PT}^2} \quad (2)$$

where

$$\delta = \left(c_{cert} - c_{PT}\right)^2 + 2\left(\sigma_{PT}^2 - \sigma_{cert}^2\right)\ln\left(\frac{\sigma_{PT}}{\sigma_{cert}}\right) \quad (3)$$

When c_1 and c_2 are known, the probability calculation is possible by the formula:

$$P = \int_{-\infty}^{c_1} f_{cert}\,dc + \int_{c_1}^{c_2} f_{PT}\,dc + \int_{c_2}^{\infty} f_{cert}\,dc =$$
$$1 + \varphi\left(\frac{c_1 - c_{cert}}{\sigma_{cert}}\right) + \varphi\left(\frac{c_2 - c_{PT}}{\sigma_{PT}}\right) - \varphi\left(\frac{c_1 - c_{PT}}{\sigma_{PT}}\right) - \varphi\left(\frac{c_1 - c_{cert}}{\sigma_{cert}}\right) \quad (4)$$

where ϕ stands for the normalized normal distribution function. For example, calculations by formulas 2–4 in the case shown in Fig. 7 yield c_1 = 12.16, c_2 = 12.58 and P = 0.58.

Information on the distributions of both PT results and CRM data is limited by experimental statistical sample sizes. Therefore, the common area P under the probability density function curves of the distributions (the probability of obtained PT results belonging to the population of the CRM data) can adequately characterize the metrological compatibility only as much as the goodness-of-fit of empirical and theoretical distributions is high. However, the P value is of practical importance since it allows one to choose a suitable null hypothesis for a criterion of a "yes–no" type for assessment of the metrological compatibility of relatively small (not infinite) number of PT results.

A-2. NULL AND ALTERNATIVE HYPOTHESES

The chosen null hypothesis H_0 states that the metrological compatibility is satisfactory if the bias $|c_{PT} - c_{cert}|$ exceeds σ_{cert} only by a value which is insignificant in comparison with random interlaboratory errors

$$H_0: \left|c_{PT} - c_{cert}\right| \le \left[\sigma_{cert}^2 + \left(0.3\sigma_{PT}\right)^2\right]^{1/2} \quad (5)$$

A coefficient of 0.3 is used according to the known metrological rule defining one standard deviation insignificant in comparison with another one when the former does not exceed 1/3 of the latter (i.e., the first variance is smaller than the second one by an order). By this hypothesis, the probability P of considering the PT results as belonging to the population of CRM data is $P \ge 0.53$ for the ratio $\gamma = \sigma_{cert}/\sigma_{PT} \ge 0.4$ (as shown in Fig. 7), when the right-hand side of expression 5 reaches the value of $1.25\sigma_{cert}$.

The alternative hypothesis H_1 assumes that the metrological compatibility is not satisfactory and the bias $|c_{PT} - c_{cert}|$ exceeds σ_{cert} significantly, for example:

$$H_1: \left|c_{PT} - c_{cert}\right| = 2.0\left[\sigma_{cert}^2 + \left(0.3\sigma_{PT}\right)^2\right]^{1/2} \tag{6}$$

etc.

A-3. CRITERION FOR NORMAL DISTRIBUTION OF PT RESULTS

The criterion for not rejecting H_0 for a statistical sample of size N, i.e., for results of N laboratories participating in the PT, is

$$\left|c_{PT/av} - c_{cert}\right| + t_{1-\alpha} s_{PT} / \sqrt{N} \le \left[\sigma_{cert}^2 + \left(0.3\sigma_{PT}\right)^2\right]^{1/2} \tag{7}$$

where $c_{PT/av}$ 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 $|c_{PT} - c_{cert}|$; $t_{1-\alpha}$ is the quantile of the one-tailed Student's distribution for the number of degrees of freedom $N - 1$; the $1 - \alpha$ value is the probability of the bias not exceeding the upper limit of its confidence interval.

By substituting the ratio γ and $s_{PT}/\sigma_{PT} = [\chi_\alpha^2/(N-1)]^{1/2}$, where χ_α^2 is the α quantile of χ^2 distribution for the number of degrees of freedom $N - 1$, into formula 7, the following transformation of the criterion is obtained:

$$\left|c_{PT/av} - c_{cert}\right| / s_{PT} \le \left[\frac{(N-1)}{\chi_\alpha^2}\left(0.09 + \gamma^2\right)\right]^{1/2} - \frac{t_{1-\alpha}}{\sqrt{N}} \tag{8}$$

Table 1 gives the numerical values for the right-hand side of the criterion at $\alpha = 0.025$.

Table 1 Bias norms in s_{PT} units by criterion 8.

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

These values are the norms for the bias of the average PT result from the analyte concentration certified in the CRM (in s_{PT} units). The value of γ should be set based on the requirements to the analytical results taking into account σ_{PT} fit-for-intended-use value that is equal either to the standard analytical/measurement uncertainty or to the target standard deviation σ_{targ} calculated using the Horwitz curve [2,3] or another database.

A-3.1 Example

According to the ASTM standard [29], the means of the results of duplicate aluminum determinations in coal fly ashes 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 aluminum. 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 0.95 quantile of the range

distribution. In case of the discussed SRM 2690 with σ_{cert} = 0.14 % the value γ is 0.14/0.38 = 0.4. Simulated statistical samples of the PT results are given in Table 2. Metrological compatibility of results of the first 15 laboratories can be assessed as satisfactory by the norm in Table 1 for γ= 0.4 (0.23), since $|c_{PT/av} - c_{cert}| = |12.30 - 12.35| = 0.05 < 0.23\ s_{PT} = 0.23 \times 0.34 = 0.08$ % as mass fraction. The same is true concerning the metrological compatibility of results of all the 30 laboratories (the norm in Table 1 is 0.30): $|c_{PT/av} - c_{cert}| = |12.38 - 12.35| = 0.03 < 0.30\ s_{PT} = 0.30 \times 0.35 = 0.11$ %.

Table 2 PT results of aluminum determination in SRM 2690 (simulated in % as mass fraction).

Lab No. i	100 c_i	Lab No. i	100 c_i
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

For other detailed examples, see Examples 3 and 4 in Annex B.

A-3.2 Reliability of the assessment

Reliability in such metrological compatibility 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 (i.e., when the alternative hypothesis H_1 is true). Criterion 8 does not allow rejecting hypothesis H_0 with probability $1 - \alpha$ 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 α. The probability of rejecting H_0, when it is false, i.e., when the alternative hypotheses H_1 are actually true (the criterion power, P_C) is

$$P_C = \varphi\left\{\frac{t_\alpha + \lambda}{\left[1 + t_{1-\alpha}^2 / 2\left(N-1\right)\right]^{1/2}}\right\} \tag{9}$$

where

$$\lambda = \frac{\left|c_{PT} - c_{cert}\right| - \sigma_{PT}\left(0.09 + \gamma^2\right)^{1/2}}{\sigma_{PT} / \sqrt{N}} \tag{10}$$

The value of the deviation parameter λ is calculated substituting the bias $|c_{PT} - c_{cert}|$ in eq. 10 by its value corresponding to the alternative hypothesis. For hypothesis H_1 by formula 6, the substitution is $2.0[\sigma^2_{cert} + (0.3\sigma_{PT})^2]^{1/2}$ and, therefore, $\lambda = [(0.09 + \gamma^2)N]^{1/2}$. The probability of an error of type II (not rejecting the H_0 when it is false) equals $\beta = 1 - P_C$. Both operational characteristics of the criterion P_C and β are shown in Fig. 8 at $\alpha = 0.025$ for different γ values and different numbers N of the PT participants.

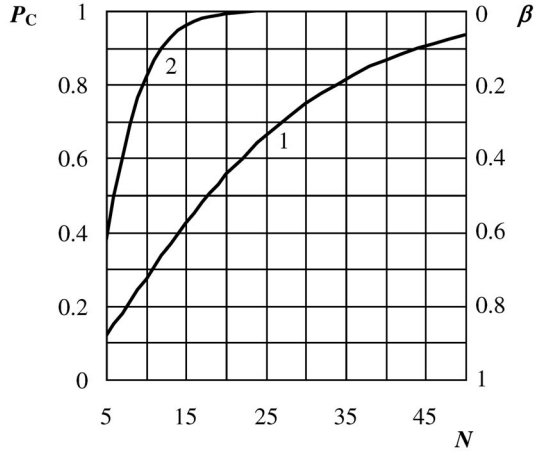

Fig. 8 Power P_C of the criterion and probability β of an error of type II (in dependence on the number N of laboratories participating in PT) for probability α = 0.025 of an error of type I; reproduced from ref. [28] by permission of Springer. Curve 1 are at γ = 0.4, and curve 2 - at γ = 1.0.

Thus, the reliability of the compatibility assessment using the hypotheses H_0 against H_1 for the PT scheme for aluminum determination in coal fly ashes (where γ = 0.4) can be characterized by (i) probability $1 - \alpha = 0.975$ of the correct assessment of the compatibility as successful (i.e., not rejecting the null hypothesis H_0 when it is true) for any number N of the laboratories participating in PT, and by (ii) probability $P_C = 0.42$ of correct assessment of the compatibility as unsuccessful (i.e., rejecting H_0 when the alternative hypothesis H_1 is true) for $N = 15$, and probability $P_C = 0.75$ for $N = 30$ results. Probability α of a type I error 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.

The power of criterion 8 is high ($P_C > 0.5$) for a number of PT participants $N \geq 20$.

A-4. NON-PARAMETRIC TEST FOR PT RESULTS WITH AN UNKNOWN DISTRIBUTION

In the case of unknown distributions differing from the normal distribution, the median is more robust than the average, i.e., it is reproduced better in the repeated experiments, being less sensitive to extreme results or outliers. Therefore, the null hypothesis, assuming here that the bias of PT results exceeds σ_{cert} by a value that is insignificant in comparison with random interlaboratory errors, has the following form:

$$H_0: \left|M_{PT} - c_{cert}\right| \leq \left[\sigma^2_{cert} + \left(0.3\sigma_{PT}\right)^2\right]^{1/2} = \Delta \tag{11}$$

where M_{PT} is the median of PT results of hypothetically infinite number N of participants, i.e., the population median.

If $M_{PT} \geq c_{cert}$, the null hypothesis H_0 implies that probability P_e of an event when a result c_i of the i-th PT-participating laboratory exceeds the value $c_{cert} + \Delta$, is $P_e\{c_i > c_{cert} + \Delta\} \leq ½$ according to the median definition. If $M_{PT} < c_{cert}$, the probability of c_i yielding the value $c_{cert} - \Delta$ is also $P_e\{c_i < c_{cert} - \Delta\} \leq ½$. The alternative hypothesis H_1 assumes that the bias exceeds σ_{cert} significantly and probabilities of the events described above are $P_e > ½$, for example:

$$H_1 : \left| M_{PT} - c_{cert} \right| = 2\Delta \tag{12}$$

where Δ is the same as in eq. 11. Probabilities P_e of the events according to the alternative hypothesis H_1 for a normal distribution (depending on the permissible bias Δ in σ_{PT} units at different γ values) are shown in Table 3.

Table 3 Probability P_e according to alternative hypothesis H_1.

γ	Δ/σ_{PT}	P_e
0.4	0.50	0.69
0.7	0.75	0.77
1.0	1.04	0.85

Since the population median is unknown in practice, and results of N laboratories participating in PT form a N-size statistical sample from the population, hypothesis H_0 is not rejected when the upper limit of the median confidence interval does not exceed $c_{cert} + \Delta$, or the lower limit does not yield $c_{cert} - \Delta$. The limits can be evaluated based on the simplest non-parametric sign test [30]. According to this test, the number N_+ of results $c_i > c_{cert} + \Delta$ or the number N_- of results $c_i < c_{cert} - \Delta$ should not exceed the critical value A (the bias norm) in order not to reject H_0. The A values are available, for example, in ref. [31]. For N from 5 to 50 PT participants and confidence levels 0.975 (α = 1–0.975 = 0.025) and 0.95 (α = 0.05), these values are shown in Table 4. The A value for fewer than six participants at α = 0.025 cannot be determined, and therefore, is not presented in Table 4 for N = 5.

Table 4 Bias norms A by the sign test.

α	N						
	5	10	15	20	30	40	50
0.025	–	1	3	5	9	13	17
0.05	0	1	3	5	10	14	18

A-4.1 Reliability of the test

The test does not allow rejecting hypothesis H_0 with a probability of $1 - \alpha$, when it is true. The probability of an error of type I by this test (to reject the H_0 hypothesis when it is true) is α. The probability of rejecting the null hypothesis when it is false, i.e., when the alternative hypothesis is actually true (the test power: P_T), is tabulated in ref. [31]. The probability of type II error (not rejecting H_0 when it is false) equals $\beta = 1 - P_T$. The operational characteristics of the test (P_T and β) are shown in Fig. 9 at α = 0.025 for the alternative hypothesis H_1 at different γ values and different numbers N of the PT participants.

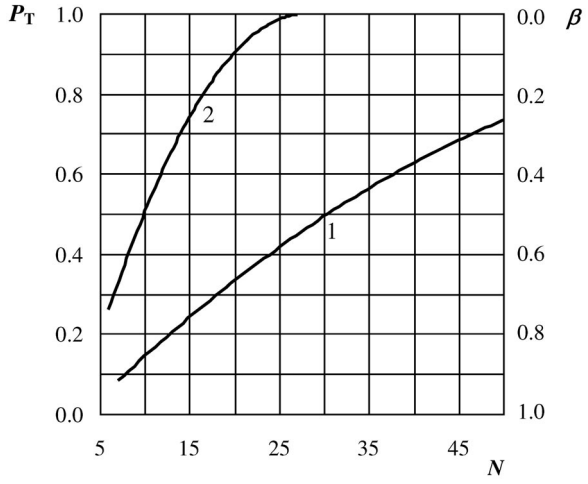

Fig. 9 Power P_T of the test and probability β of an error of type II in dependence on the number N of laboratories participating in PT, when probability of an error of type I is $\alpha = 0.025$; reproduced from ref. [30] by permission of Springer. The null hypothesis H_0 is tested against the alternative hypotheses H_1 at $\gamma = 0.4$ and $\gamma = 1.0$ shown by curves 1 and 2, respectively.

A-4.2 Example

The hypothesis about normal distribution of the PT results in the example shown in Table 2 was not tested because of the small size of the statistical samples. Therefore, the sample size is increased here to $N = 50$: the simulated data are presented in Table 5 (the simulation is performed by the known method of successive approximations). Such sample size allows testing the hypothesis about the data normal distribution applying the Cramer-von-Mises ω^2-criterion, which is powerful for statistical samples of small sizes [32]

$$\omega^2 = -N - 2\sum_{j=1}^{N}\left\{\left[\left(2j-1\right)/2N\right]\ln\phi\left(x_j\right) + \left[1-\left(2j-1\right)/2N\right]\ln\left[1-\phi\left(x_j\right)\right]\right\} \tag{13}$$

where $j = 1, 2, \ldots, N$ is the number of the PT result C_j in the statistical sample ranked by increasing c value ($c_1 \le c_2 \le \ldots \le c_N$); $x_j = (c_j - c_{PT/av})/s_{PT}$ is the normalized value of the j-th result which is distributed with the mean of 0 and the standard deviation of 1; and $\phi(x_j)$ is the value of the function of the normalized normal distribution for x_j.

Table 5 PT results of aluminum determination in SRM 2690 (simulated in % as mass fraction) ranked according to their increasing value.

No. j	Result, 100 C_i	100 $(C_j - c_{cert})$	Sign	No. j	Result, 100 C_i	100 $(C_j - c_{cert})$	Sign	No. j	Result, 100 C_i	100 $(C_j - c_{cert})$	Sign
1	11.86	–0.49	–	18	12.44	0.09	0	35	12.53	0.18	0
2	11.88	–0.47	–	19	12.44	0.09	0	36	12.55	0.20	+
3	11.90	–0.45	–	20	12.45	0.10	0	37	12.56	0.21	+
4	11.91	–0.44	–	21	12.46	0.11	0	38	12.57	0.22	+
5	11.93	–0.42	–	22	12.46	0.11	0	39	12.60	0.25	+
6	11.96	–0.39	–	23	12.47	0.12	0	40	12.61	0.26	+
7	11.96	–0.39	–	24	12.48	0.13	0	41	12.64	0.29	+
8	11.97	–0.38	–	25	12.49	0.14	0	42	12.66	0.31	+
9	11.98	–0.37	–	26	12.49	0.14	0	43	12.67	0.32	+
10	11.99	–0.36	–	27	12.50	0.15	0	44	12.68	0.33	+
11	12.03	–0.32	–	28	12.50	0.15	0	45	12.69	0.34	+
12	12.07	–0.28	–	29	12.51	0.16	0	46	12.76	0.41	+
13	12.17	–0.18	0	30	12.51	0.16	0	47	12.81	0.46	+
14	12.19	–0.16	0	31	12.52	0.17	0	48	12.84	0.49	+
15	12.20	–0.15	0	32	12.52	0.17	0	49	12.90	0.55	+
16	12.34	–0.01	0	33	12.53	0.18	0	50	12.96	0.61	+
17	12.43	0.08	0	34	12.53	0.18	0	$N_- = 12$;		$N_+ = 15$	

The probability that $\omega^2 = 1.95$ calculated by formula 13 for the data in Table 5 exceeds randomly the critical value 1.94 (for $N = 50$) equals 0.10 [31]. Therefore, the hypothesis about normal distribution of these data should be rejected at the level of confidence of 0.90. The corresponding empirical histogram and the theoretical (normal) distribution are shown in Fig. 10. It is clear that the empirical distribution is bimodal, therefore, no normal distribution can fit it. Since other known distributions are also not suitable here, the proposed non-parametric test was applied for the comparability assessment of the results.

Taking into account $c_{cert} = 12.35$ %, $\sigma_{cert} = 0.14$ %, $\sigma_{PT} = 0.38$ %, and $\gamma = 0.14/0.38 = 0.4$, one can calculate $\Delta = 0.50 \times 0.38 = 0.19$ % (Table 5), $c_{cert} + \Delta = 12.54$ % and $c_{cert} - \Delta = 12.16$ %. There are $N_+ = 15$ results $c_j > 12.54$ %, $N_- = 12$ results $c_j < 12.16$ %, and $N - N_+ - N_- = 23$ values in the range $c_{cert} \pm \Delta$. The sample median found is $c_{25} = c_{26} = 12.49 > c_{cert} = 12.35$ % and $N_+ > N_-$. However, N_+ is lower than the critical value $A = 17$ at $\alpha = 0.025$ and $N = 50$ (Table 4). Therefore, null hypothesis H_0 concerning successful metrological compatibility of the results is not rejected.

Reliability of the assessment with hypotheses H_0 against H_1 for this case can be characterized by: (i) probability $1 - \alpha = 0.975$ of correct assessment of the compatibility as successful (not rejecting the null hypothesis when it is true) for any number $N \geq 6$ of the PT participants, and (ii) probability $P_T = 0.73$ of correct assessment of the compatibility of $N = 50$ PT results as unsuccessful (rejecting H_0 when alternative hypothesis H_1 is true). The probability α of a type I error is 0.025 for any $N \geq 6$, while the probability β of a type II error is 0.27 for $N = 50$.

Additional examples of the use of the sign test see in Annex B, Examples 1 and 2, of ω^2-criterion application: Example 3.

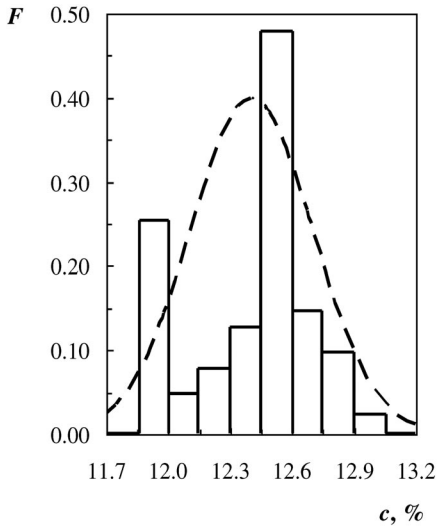

Fig. 10 Histogram of PT results (frequency F of a result value c) – solid line, and the fitted normal distribution – dotted line; reproduced from ref. [30] by permission of Springer.

A-4.3 Limitations

Since the critical A values from the sign test are determined for $N \geq 4$–8 depending on probabilities α, and the test power is calculated also only for $N \geq 6$–8, the proposed metrological compatibility assessment cannot be performed for a smaller sample size. 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. For example, practically the same power (0.73 and 0.75) was achieved in the sign test of the compatibility of PT results for aluminum determination in coal fly ashes at $N = 50$ discussed above, and in the t-test for the same purpose at $N = 30$ in the previous paragraph 3. The power efficiency here is approximately of 30/50 = 0.6. On the other hand, when information about the distribution of PT results is limited by $N < 50$, it is a problem to evaluate the goodness-of-fit empirical and theoretical/normal distributions, a decrease of the t-test power and the corresponding decrease of reliability of the compatibility assessment caused by deviation of the empirical distribution from the normal one.

ANNEX B. EXAMPLES

CONTENTS

EXAMPLE 1. SCENARIO 1: PT FOR LEAD DETERMINATION IN AIRBORNE PARTICLES

B-1.1 Objectives of the PT

The objectives of this PT were to determine whether the quality criteria described in the European Directives [33,34] concerning the analysis of As, Cd, Ni, and Pb in airborne particles are reached and the most important sources of uncertainties are identified. The measurement method is divided by the standard [35] into two main parts: first, the sampling in the field, and, second, the analysis in the laboratory. During sampling, particles are collected by drawing a measured volume of air through a filter mounted in a sampler designed to collect the fraction of suspended particulate matter of less than 10 μm (PM10) [36]. The sample filter is transported to the laboratory, and the analytes are taken into solution by closed-vessel microwave digestion using nitric acid and hydrogen peroxide. The resultant solution is analyzed by known analytical methods. When the quantity of an analyte in the solution is measured, its concentration can be expressed in ng/m^3 of the sampled air.

The PT was organized in 2005 in France and focused on the second (analytical) part of the method. The PT provider was the Ecole des Mines de Douai (EMD) supported by the Laboratoire National de Métrologie et d'Essais (LNE). Ten laboratories (N = 10) of the Association Agréées de Surveillance de la Qualité de l'Air participated in this trial.

Only results for lead are discussed below for brevity.

B-1.2 Procedure for preparation of the IHRM

The PM10 fraction of suspended particulate matter was collected by EMD on an industrial site according to the standard [36]. Sampling was performed on 20 quartz filters (diameter 50 mm) during one week at a flow rate of air of 1 m^3 h^{-1}, which means a total of 168 m^3. Dust on the filters was then digested with 5 ml HNO_3 + 1 ml H_2O_2 in a microwave oven.

The LNE was in charge to prepare 1 l of a solution from the digestion residue, which could be used in the PT as an IHRM. The assigned/certified value of the lead content in the solution c_{cert} = 26.72 μg l^{-1} provided by LNE was obtained with a primary method: isotope dilution-inductive coupled plasma-mass spectrometry (ID-ICP-MS). This content corresponds to 26.72 × 1000/168 = 159 ng m^{-3} Pb in the sampled air. The expanded measurement uncertainty of the certified value was U_{cert} = 0.77 μg l^{-1} at the level of confidence 0.95 and the coverage factor of 2. No stability tests were conducted, since the laboratories used the solution just after the preparation. The uncertainty due to inhomogeneity of the 1-l solution was considered negligible. Note that the standard uncertainty was u_{cert} = 0.77/2 = 0.38 μg l^{-1}, i.e., 1.4 % of the certified value.

Each laboratory received a bottle of 50 ml of this solution (for all analytes).

B-1.3 Analytical methods used and raw data

The list of the participating laboratories was confidential. All of them followed the standard [35]. The methods used were: ICP-MS, graphite furnace atomic absorption spectrometry (GF-AAS), and inductively coupled plasma optical emission spectroscopy with ultrasonic nebulization (ICP-OES-USN). The measurements results of *i*-th laboratory c_i, i = 1, 2, …, N = 10, are shown in Table 6.

Table 6 Results of the PT for lead content determination in the solution.

Lab No. *i*	Method	c_i/μg l^{-1}	$(c_i - c_{cert})$/μg l^{-1}	z_i	Sign
1	ICP-MS	20.12	–6.60	–1.98	–
2	ICP-MS	20.28	–6.44	–1.93	–
3	ICP-OES-USN	30.34	3.62	1.08	+
4	GF-AAS	29.00	2.28	0.68	+
5	ICP-MS	25.00	–1.72	–0.51	–
6	GF-AAS	28.40	1.68	0.50	+
7	ICP-MS	27.80	1.08	0.32	+
8	ICP-MS	25.70	–1.02	–0.31	–
9	GF-AAS	28.20	1.48	0.44	+
10	ICP-MS	25.51	–1.21	–0.36	–

B-1.4 Statistical analysis of the data

There was no statistically significant dependence of the results on the analytical method used. The robust value of the experimental standard deviation s_{PT} of a laboratory result c_i calculated by the LNE from the data shown in Table 6 using Algorithm A of the standards [3,37] was of 3.93 μg l^{-1}, i.e., 14.7 % of the certified value. Since the expanded uncertainty stated for lead in the European Directives [33,34] and the standard [35, p. 30] is 25 %, the target value for standard deviation of a laboratory result in the PT was σ_{targ} = 25/2 = 12.5 % or 3.34 μg l^{-1}.

The uncertainty of the certified value u_{cert} = 1.4 % was negligible in comparison with σ_{targ} and the z-score was applicable for proficiency testing based on the target σ_{targ} value. The calculated z-score values are shown in Table 6. All of them are between –2 and +2, and therefore, were interpreted as satisfactory.

B-1.4.1 Metrological compatibility assessment

Since a hypothesis on the normal distribution of the PT results was not taken into account, compatibility of the results (as a group) is tested based on non-parametric statistics as shown in Annex A, paragraph 4.

As the standard uncertainty of the certified value u_{cert} = 1.4 % was insignificant in comparison with the target standard deviation of PT results σ_{targ} = 12.5 %, the permissible bias of the median of the PT results from the certified value was $\Delta = 0.3\sigma_{targ}$ = 3.75 % or 1.00 μg l^{-1}. Therefore, $c_{cert} + \Delta$ = 27.72 μg l^{-1} and $c_{cert} - \Delta$ = 25.72 μg l^{-1}. There were $N_+ = 5$ results $c_i > 27.72$ μg l^{-1} and $N_- = 5$ results $c_i <$ 25.72 μg l^{-1}. They are shown in Table 6 as signs "+" and "–", respectively. Both N_+ and N_- values are higher than the critical value $A = 1$ in Table 4. Therefore, null hypothesis H_0 concerning compatibility of this group of results should be rejected, in spite of the satisfactory z-score values for every laboratory participant of the PT. Probability of a type I error (to reject the hypothesis when it is correct) of the decision is of 0.025, while probability of a type II error (to not reject the hypothesis when it is false) is of above 0.85 according to Fig. 9.

EXAMPLE 2. SCENARIO 2: PT FOR ARSENIC DETERMINATION IN WATER

B-2.1 Objectives of the PT

The aim of the PT was to support water-testing laboratories from the Southern African Development Community (SADC) and from the East African Community in their effort to improve the quality of measurement results. The PT round was organized in 2006 within the Water PT Scheme of the SADCMET (SADC Cooperation in Measurement Traceability). The organizers were the Water Quality Services, Windhoek, Namibia, in cooperation with the Universität Stuttgart, Germany, and with financial support by the Physikalisch-Technische Bundesanstalt, Braunschweig, Germany. The analytes were Ca, Mg, Na, K, Fe, Mn, Al, Pb, Cu, Zn, Cr, Ni, Cd, As, SO_4^{2-}, Cl^-, F^-, NO_3^-, and PO_4^{3-} in synthetic water modeling drinking/ground water. Three IHRMs with different analyte concentrations were prepared and distributed between the laboratory participants for analysis.

In the following description, the determination of the arsenic concentration in one IHRM only was selected as an example.

B-2.2 Procedure for preparation of the IHRM

The IHRM was formulated on the basis of analytical-grade water spiked with pure chemicals. Arsenic(III) oxide from Sigma-Aldrich (purity p_c = 99.995 %) was used for the preparation of the stock solution with a content of As of about 0.4 mg g^{-1}. The mass $m_{As_2O_3}$ of the oxide was measured on an analytical balance (Sartorius RC 210D), the total mass $m_{ss/t}$ of the stock solution was determined by the difference weighing on a Sartorius BA3100P balance. About m_{ss} = 100 g of the stock solution was diluted to about $m_{dil/t}$ = 1000 g also on a Sartorius BA3100P balance. Finally, about m_{dil} = 200 g of the diluted solution (also weighed on the same balance) was diluted to about m_{lot} = 49 900 g. The total mass m_{lot} of this lot was determined by difference weighing on a Sartorius F150S balance.

The assigned/certified value of the As concentration in the IHRM was assessed according to the preparation procedure and taking into account the proportion $p_{As/As_2O_3} = M(As)/M(As_2O_3)$, where M is the atomic or molecular weight (from IUPAC publications), the purity p_c of As_2O_3 used, the density ρ_{lot} of the final lot, and a buoyancy correction factor b_{cf}. The density of the final lot was measured gravimetrically using a 100-ml pycnometer. The certified value c_{cert} of the mass concentration of As in the final lot was calculated by the following formula:

$$c_{cert} = \frac{m_{As_2O_3} \cdot p_{As/As_2O_3} \cdot p_c \cdot m_{ss} \cdot \rho_{lot} \cdot m_{dil}}{m_{ss/t} \cdot m_{lot} \cdot b_{cf} \cdot m_{dil/t}} \quad (14)$$

Formula 14 enables also calculation of the uncertainty budget of the certified value. The uncertainties of the masses were derived from precision experiments, delivering directly the standard uncertainty, and from the linearity tolerances given by the manufacturer (used as rectangular distribution). The uncer-

tainty of the purity was derived from the manufacturer's information. The uncertainty of the buoyancy correction factor was estimated from the possible variations in the atmospheric pressure, air humidity and temperature [38]. For the estimation of the uncertainty of density, a separate budget was calculated taking into account the uncertainties of the weighing and that of the temperature measurement. The uncertainties of the atomic weights and of stability and homogeneity of the solution were neglected.

The assigned/certified value of the As content in the IHRM and its expanded uncertainty were $c_{cert} \pm U_{cert} = 0.1706 \pm 0.0001$ mg l^{-1} at the level of confidence 0.95 and the coverage factor of 2. Note that the expanded uncertainty was of 0.07 % of the reference value.

Each laboratory received a bottle of 1 l of this IHRM (for all analytes).

B-2.3 Analytical methods used and raw data

Nine laboratory participants ($N = 9$) reported results on determination of the As concentration shown in Table 7. One of the current major problems with water analysis in Africa is absence of any common standard for analytical methods. The methods used were ICP-OES, AAS, and others.

Table 7 Results of the PT for arsenic content determination in water.

Lab No. i	Method	c_i/mg l^{-1}	$(c_i - c_{cert})$/mg l^{-1}	z_i	Comment	Sign
4	AAS	0.03	–0.1406	–4.12	No	–
10	other	0.20	0.0294	0.86	Yes	+
18	ICP-OES	0.20	0.0294	0.86	Yes	+
19	ICP-OES	0.12	–0.0506	–1.48	Yes	–
26	ICP-OES	0.12	–0.0506	–1.48	Yes	–
34	AAS	0.169	–0.0206	–0.05	Yes	0
35	AAS	0.08	–0.0906	–2.66	Quest	–
37	ICP-OES	0.789	0.6184	18.12	No	+
38	other	0.258	0.0874	2.56	Quest	+

B-2.4 Statistical analysis of the data

High standard deviations from the certified value (above 20 % of the value) were expected at a workshop organized for representatives of the laboratory participants prior to this PT round. Therefore, it was decided to use the target standard deviation σ_{targ} of 20 % of the certified value, when the experimental standard deviation $s_{PT} > 20$ %. Since in the As case the robust s_{PT} value, calculated from the data shown in Table 7 by Algorithm A of the standards [3,37], was of 50.5 % (0.086 mg l^{-1}), the stated target value $\sigma_{targ} = 20$ % (0.034 mg l^{-1}) was applied for the proficiency assessment with z-score. The z-score values are shown in Table 7 with the comments: satisfactory (Yes) when they were between –2 and +2, questionable (Quest) for $2 < |z| < 3$, and not satisfactory (No) for the other values. Thus, results of five laboratories were assessed as satisfactory, of two – as questionable, and of two – as not satisfactory.

There was no statistically significant dependence of the results on the analytical methods used.

B-2.4.1 Metrological compatibility assessment

Since a hypothesis on the normal distribution of the PT results was not taken into account, compatibility of the results (as a group) is tested based on non-parametric statistics as shown in Annex A, paragraph 4. The standard uncertainty of the certified value $u_{cert} = 0.07/2 = 0.035$ % was insignificant in comparison with the target standard deviation of PT results $\sigma_{targ} = 20$ %.

Therefore, the permissible bias of the median of the PT results from the certified value was $\Delta = 0.3\sigma_{targ} = 6$ % or 0.0102 mg l^{-1}, $c_{cert} + \Delta = 0.1808$ mg l^{-1} and $c_{cert} - \Delta = 0.1604$ mg l^{-1}. There were

$N_+ = 4$ results $c_i > 0.1808$ mg l^{-1}, $N_- = 4$ results $c_i < 0.1604$ mg l^{-1}, and $N - N_+ - N_- = 1$ result in the range $c_{cert} \pm \Delta$. They are shown in Table 7 as signs "+", "–", and "0", respectively. Both N_+ and N_- values were higher than the critical value $A = 1$ in Table 4. Therefore, null hypothesis H_0 concerning compatibility of this group of the PT results was rejected with probability of a type I error (to reject the hypothesis when it is correct) of 0.025 and probability of a type II error (to not reject the hypothesis when it is false) of above 0.85 according to Fig. 9.

EXAMPLE 3. SCENARIO 3: PT FOR DETERMINATION OF CONCRETE COMPRESSIVE STRENGTH

B-3.1 Objectives of the PT

The purpose of the PT organized in 2005 by the Israel Laboratory Accreditation Authority and the National Physical Laboratory of Israel (INPL) was assessment of performance of accredited laboratories testing concretes for a local (Israeli) building industry. Slump and compressive strength were chosen in the PT as the test parameters of fresh and hardened concrete, practically the most required by the customers.

The PT was based on preparing and immediate use of test items/samples of an IHRM of a concrete at a reference laboratory (RL)—the Research Unit of the Department of Building Units and Materials at ISOTOP, Ltd.

The results of $N = 25$ PT participants were compared with the IHRM assigned/certified values, taking into account both the uncertainties of the certified values and the measurement/test uncertainties of the participants [39].

In the following description, the compressive strength results only are discussed as an example. Compressive strength is measured by standard [40] as a pressure, MPa, applied by a special testing machine to a 100-mm hardened concrete test cube in order to destroy it. To prepare the test cubes, the standard requires filling the corresponding steel forms with the concrete by hand using a steel rod or by means of a vibrating table. Afterwards, the test cubes should be stored 7 days under controlled conditions (air temperature of 21 ± 3 °C and humidity of more than 95 %) and then 21 days under standard laboratory conditions for hardening. On the 28th day, the test cubes should be destroyed. Six test cubes are recommended to be prepared and destroyed as replicates. A test result is calculated as an average of the six pressure measurements.

B-3.2 Procedure for preparation of the IHRM

The composition of the IHRM developed for the PT corresponded to fresh concrete of type B30 by the standard [41]. Aggregates were thoroughly washed with water before the experiment, dried until constant weight at 105 ± 5 °C, sieved and homogenized. The sea sand was also dried until constant weight at 105 ± 5 °C, sieved (the fraction smaller than 0.65 mm was used) and homogenized. The components were stored in RL at air humidity of 45 to 60 %.

The concrete for every PT participant (IHRM sample of 35 l) was produced by RL using the same Pan Mixer of 55 l, company "Controls", Italy, in the same conditions. Every participant had a possibility to start testing its sample from the moment when the concrete preparation was finished.

Twenty-nine samples were prepared by RL during two weeks in September 2005 before the rainy season in Israel influences air humidity. RL tested four samples—the 1st, 12th, 23rd, and the 29th (last) samples—for the material inhomogeneity study and characterization. Other 25 samples were tested by the PT participants according to the schedule preliminary prepared and announced.

B-3.2.1 IHRM homogeneity, certified value, and its uncertainty

RL prepared the test cubes from the IHRM samples by filling the corresponding steel forms with a vibrating table. Since the samples were tested immediately after preparation, its stability was not relevant as a source of measurement/test uncertainty. Between- and intra- sample inhomogeneity was evaluated based on analysis of variances (ANOVA).

The assigned/certified value of the IHRM was calculated as averaged RL result of the compressive strength determination in the four samples: $c_{\text{cert}} = \sum_{n=1}^{4} c_{\text{avg}/n}/4 = 32.0$ MPa, where $c_{\text{avg}/n} = \sum_{j=1}^{6} c_{nj}/6$ is the result of the test of the n-th sample, c_{nj} is the replicate j value for the sample n.

The standard uncertainty of the certified value u_{cert} included the measurement/test uncertainty component $u_{\text{mRL}} = 1$ MPa and the components arising from the material inhomogeneity: $u_{\text{cert}} = (u_{\text{mRL}}{}^2 + s_{\text{bs}i}{}^2 + s_{isi}{}^2/6)^{1/2} = 1.9$ MPa, where $s_{\text{bs}i} = 1.53$ MPa and $s_{isi} = 0.70$ MPa are the between- and intra-sample standard deviations, respectively.

B-3.3 Methods used and raw data

The participants prepared test cubes using their own facilities both for hand preparation ("hand" in Table 8) and with a vibrating table ("vibr.") corresponding to the standard [40]. On the next day after preparation, the hardened cubes were transferred from RL to the laboratory of the participant, where compressive strength determinations were performed (every one of 6 replicates). The participant results are presented in Table 8, where $c_i = \sum_{j=1}^{6} c_{nj}/6$ is the result of the i-th laboratory; c_{ij} is the j-th replicate of the i-th laboratory; u_{mLP} is the standard measurement/test uncertainty declared by the participant; $u_{\text{comb}} = (u_{\text{mLP}}{}^2 + u_{\text{cert}}{}^2)^{1/2}$ is the test combined standard uncertainty; and $\zeta_i = (c_i - c_{\text{cert}})/u_{\text{comb}}$ is the ζ-score value for the i-th laboratory result.

Table 8 Results of compressive strength determination obtained by the PT participants.

Lab No. i	Cube prep.	Replicates, c_{ij}/MPa						u_{mLP}/ MPa	c_i/ MPa	$(c_i - c_{\text{cert}})$/ MPa	u_{comb}/ MPa	ζ_i
		$c_{\text{i}1}$	$c_{\text{i}2}$	$c_{\text{i}3}$	$c_{\text{i}4}$	$c_{\text{i}5}$	$c_{\text{i}6}$					
1	hand	27.5	27.5	26	28.5	29	28	1.9	27.75	–4.2	2.6	–1.60
2	vibr	30.5	29	30.5	28.5	29	30	5.2	29.58	–2.4	5.5	–0.43
3	hand	33	33	33	33	33	33	1.9	33.00	1.0	2.6	0.39
4	hand	31.5	31.5	32	30	31	32	1	31.33	–0.6	2.1	–0.31
5	hand	30.5	31	30	30.5	31.5	30	0.67	30.58	–1.4	2.0	–0.71
6	hand	29.5	30	29	29	30	28.5	5.2	29.33	–2.6	5.5	–0.48
7	hand	29	27	29.5	31.5	27.5	29.5	2.5	29.00	–3.0	3.1	–0.97
8	hand	27.5	27	26.5	27.5	27.5	27.5	2	27.25	–4.7	2.7	–1.74
9	hand	30.5	30	30.5	29	30.5	29.5	1.9	30.00	–2.0	2.6	–0.75
10	vibr	34	33	34	33	33.5	33	0.41	33.42	1.4	1.9	0.76
11	hand	30	29.5	28.5	30.5	29	30.5	0.67	29.67	–2.3	2.0	–1.17
12	hand	31	31	31	30.5	31	30	5.2	30.75	–1.2	5.5	–0.22
13	hand	28	27.5	27	28	28.5	29	1	28.00	–4.0	2.1	–1.89
14	hand	32	30.5	31.5	32	30.5	31	1	31.25	–0.7	2.1	–0.35
15	hand	33.5	33	33	33	32.5	32	5.2	32.83	0.9	5.5	0.15
16	hand	26.5	20.5	27	27	27.5	27.5	1.9	26.00	–6.0	2.6	–2.26
17	hand	31	30	29.5	28.5	29.5	29.5	1	29.67	–2.3	2.1	–1.10
18	hand	31	30	30	31	30.5	30	1.9	30.42	–1.6	2.6	–0.59
19	hand	30	30	29.5	28.5	29.5	28.5	2.5	29.33	–2.6	3.1	–0.85
20	hand	32.5	32.5	31	32	31	31.5	5.2	31.75	–0.2	5.5	–0.04
21	hand	30	30	29.5	30.5	30	30.5	5.2	30.08	–1.9	5.5	–0.34

(*continues on next page*)

Table 8 (*Continued*).

Lab No. i	Cube prep.	Replicates, c_{ij}/MPa						u_{mLP}/ MPa	c_i/ MPa	$(c_i - c_{cert})$/ MPa	u_{comb}/ MPa	ζ_i
		c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}	c_{i6}					
22	hand	27	28	28	28.5	27.5	28	2.4	27.83	–4.1	3.0	–1.38
23	hand	32	31.5	31.5	32	32	31	5.2	31.67	–0.3	5.5	–0.06
24	hand	32	31	31	31.5	31	31	5.2	31.25	–0.7	5.5	–0.13
25	hand	30.5	33	31.5	33	31.5	33.5	5.2	32.17	0.2	5.5	0.03

B-3.4 Statistical analysis of the data

The hypothesis about normal distribution of all the replicates c_{ij} obtained in the PT was not rejected according to the Cramer-von-Mises criterion: the empirical value $\omega^2 = 1.53$ (calculated by formula 13, Annex A) is less than the critical value of 2.50 at the level of confidence of 0.95. The histogram of c_{ij} values and the fitted normal distribution are presented in Fig. 11.

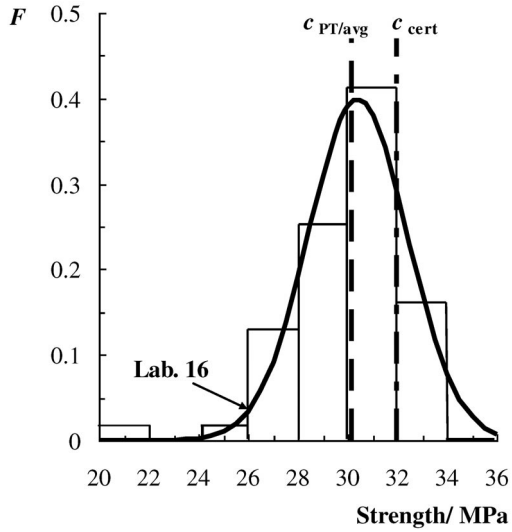

Fig. 11 Histogram (frequency F of a strength value c) and fitted normal distribution of compressive strength determination results; reproduced from ref. [39] by permission of Springer**.** The value c_{cert} is the IHRM assigned/certified strength value, and $c_{PT/avg}$ is the average strength value obtained by the PT participants. The pointer shows the average slump result obtained by laboratory (participant) No. 16.

When the hypothesis about normal distribution of replicates c_{ij} is not rejected, there are no reasons also to reject the hypothesis about normal distribution of the replicate averages c_i. The total average result of the participants $c_{PT/avg} = \sum_{i=1}^{N} c_i/25 = 30.2$ MPa is shown in Fig. 11 by a dotted line. The standard deviation of c_i from $c_{PT/avg}$ was $s_{PT} = 1.9$ MPa. The assigned/certified value $c_{cert} = 32.0$ MPa is shown in Fig. 11 by another dotted line.

All the results were satisfactory according to the ζ-score values. The only questionable score value was $2 < |\zeta = -2.26| < 3$ obtained by laboratory No. 16 ($i = 16$), shown in Fig. 11 by a pointer. Probably it was a random deviation: five laboratories out of 100, i.e., one out of 20 laboratories, can have $|\zeta| > 2$ at the level of confidence of 0.95.

B-3.4.1 Metrological compatibility assessment

Since the hypothesis on the normal distribution of the PT results was not rejected, compatibility of the results (as a group) is tested as shown in Annex A, paragraph 3.

Calibrated testing machines used in RL and in the laboratories participated in PT allow to measure pressure with standard uncertainty of less than 2 %. Therefore, the values u_{cert} and s_{PT} are equal and the assumption $\gamma = 1$ is reasonable here. Since $|c_{PT/avg} - c_{cert}|/s_{PT} = 0.95 < 1.04$, null hypothesis H_0 concerning compatibility of this group of the PT results was not rejected with probability of a type I error (to reject the hypothesis when it is correct) of 0.025 and probability of a type II error (to not reject the hypothesis when it is false) of 0.35 according to Fig. 8. Thus, assessment of the PT results based on the individual score values and on evaluation of the compatibility of these results as a group coincided for the strength determinations. Nevertheless, metrological comparability in such a case should be discussed further.

The results of measuring the pressure (applied to a test cube in order to destroy it) are traceable to the corresponding international measurement standards. The stages of test cube hardening during 28 days can be performed under conditions controlled by using traceable measurements. Therefore, traceability of the IHRM-assigned strength value to international measurement standards can be achieved theoretically. However, again, the test cube preparation depends on the technician's art, even when a vibrating table is used. Moreover, the fresh concrete is not stable and any IHRM, as described above, can be intended for the single use only and locally, where it was prepared. Thus, metrological comparability for the strength determinations is also of local relevance.

EXAMPLE 4. LIMITED POPULATION OF PT PARTICIPANTS: PT FOR ACID NUMBER DETERMINATION IN USED MOTOR OILS

B-4.1 Objectives of the PT

Acid number (A_N) of motor oil is an important parameter of its quality characterizing the aging processes in the oils, corrosive properties, etc. The oil A_N (also called "neutralization number") is defined as the mass of KOH necessary for the neutralization of acid constituents in 1 g of the oil, with the unit mg g^{-1} [42–45].

The standard methods for A_N determination in petroleum products are based on color indicator titration [42,43] or on potentiometric titration [44,45] of the oil. The specific problem of used motor oils constitutes a matrix effect. The matrix includes a number of contaminants accumulated during use of the oil that may be considered to have acidic characteristics: organic and inorganic acids, esters, phenolic compounds, lactones, resins, metal and ammonium salts, salts of other weak bases, acid salts of polybasic acids, and addition agents such as inhibitors and detergents. These contaminants lead to the dark color of used oil, and therefore, distortions complicate registration of the end-point of the titration by standard methods using a color indicator. Standard potentiometric titration methods allow analysis of such oils, but are complex and labor-consuming since they require preparation of nonaqueous buffer solutions. The general drawback of the standard methods is the use of toxic toluene, as solvent, and nonaqueous (alcoholic) titrants sensitive to atmospheric CO_2. Therefore, a pH-metric method without titration was developed at INPL [46]. However, in this method, the matrix effect has probably also an influence when the contaminants poison the glass electrode and disturb its function.

The purpose of the PT, organized in 1999–2000 by the Israel Forum of Managers of Oil Laboratories with participation of INPL, was to assess results of A_N determination in used motor oils, obtained in different laboratories by different methods [47]. There were $N_p = 12$ laboratories which had experience in the field and could participate in the PT. In the following description, the PT-2000 only is discussed as an example. Ten laboratories ($N = 10$) took part in this PT.

B-4.2 Procedure for preparation of the IHRM

About 20 l of typical used motor oil was filtered through a filter paper, heated to 70 °C, thoroughly stirred at this temperature for 7 h, and then (still hot) divided into samples stored in plastic bottles of 1 l with a hermetic cover.

Four random samples were tested for kinematic viscosity at 40 °C by the standard method [48]. The sample standard deviation of 1 % from the mean test result (90.6 cSt or 0.0906 $cm^2\ s^{-1}$) was accepted as an indicator of the oil homogeneity.

Stability of the IHRM was not studied since there are no reasons for A_N change in such an oil during a time necessary for the PT (about month). Thus, uncertainties caused by homogeneity and stability of the IHRM were neglected.

B-4.2.1 Characterization of the IHRM

Characterization of the IHRM was based on results of the potentiometric titration method obtained in the PT, since A_N is defined in terms of this method according to the standard [45]. In other words, potentiometric titration is the primary method for A_N determination by definition.

A sample of 1 l was sent to every participating laboratory. The results are shown in Table 9.

Table 9 Results of the PT for A_N determination in used motor oil.

Lab No. *i*	Method	c_i/mg g^{-1}	$(c_i - c_{cert})$/mg g^{-1}	z_i	Comment
1	Pot. titr.	3.48	0.87	2.35	Quest
2	Pot. titr.	2.26	–0.35	–0.95	Yes
3	Pot. titr.	3.28	0.67	1.81	Yes
4	Pot. titr.	3.28	0.67	1.81	Yes
6	Pot. titr.	1.17	–1.44	–3.89	No
7	Pot. titr.	2.80	0.19	0.51	Yes
8	Pot. titr.	2.42	–0.19	–0.51	Yes
10	Pot. titr.	2.39	–0.22	–0.59	Yes
11	pH-metr.	2.70	0.09	0.24	Yes
12	Pot. titr.	2.45	–0.16	–0.43	Yes

Supposing a normal distribution of potentiometric titration results ("Pot. titr." in the table), the IHRM assigned/certified value was calculated as the sample average of these results: $c_{cert} \sum_{i=1}^{10,12} c_i/9 = 2.61$ mg g^{-1}. Taking into account the effect of the limited population of the PT participants (paragraph 6 of the Guide), the standard uncertainty of the certified value was calculated as the standard deviation of the sample average: $u_{cert} = s[(N_p - N^*)/(N_p N^*)]^{1/2} = 0.71[(12 - 9)/(12 \times 9)]^{1/2} = 0.12$ mg g^{-1}, where $N^* = 9$ is the number of the potentiometric titration results, and $s = 0.71$ mg g^{-1} is their sample standard deviation from c_{cert}.

B-4.3 Methods used and raw data

In addition to the results obtained by the standard potentiometric titration method and discussed above, one can find in Table 9 the INPL result (lab No. 11) obtained with the pH-metric method without titration ("pH-metr" in the table). Any inter-method difference of the results was not indicated. The total average PT result was $c_{PT/avg} = 2.62$ with the standard deviation $s_{PT} = 0.67$ mg g^{-1}.

B-4.4 Statistical analysis of the data

Requirements of the standard [45] to reproducibility of the results are formulated as a permissible difference between two single and independent results obtained by different operators working in different laboratories on identical test material, i.e., as a range R_{A_N}. For manual titration of used lubricating oils, R_{A_N} = 39 %, and for automatic titration, R_{A_N} = 44 % of the average result $c_{PT/avg}$. Therefore, the target standard deviation can be formulated as $\sigma_{targ} = R_{A_N}/2.77$ at level of confidence 0.95 [49]. For manual titration $\sigma_{targ} = 39 \times 2.62/(2.77 \times 100) = 0.37$, and $\sigma_{targ} = 44 \times 2.62/(2.77 \times 100) = 0.42$ mg g^{-1} for automatic titration. Since even for minimal of these two values the ratio $\gamma = u_{cert}/\sigma_{targ} = 0.12/0.37 = 0.3$, z-score can be applied for the proficiency assessment. The z-score values at $\sigma_{targ} = 0.37$ are shown in Table 9 with the comments: satisfactory (Yes) when they were between –2 and +2, questionable (Quest) for $2 < |z| < 3$, and not satisfactory (No) for the other values. Thus, results of eight laboratories were assessed as satisfactory, of one – as questionable, and of another one – as not satisfactory.

B-4.4.1 Metrological compatibility assessment

Since the hypothesis on the normal distribution of the PT results was not rejected, compatibility of the results (as a group) is tested as shown in Annex A, paragraph 3.

The bias $|c_{PT/avg} - c_{cert}|/s_{PT}$ is $0.01 < 0.20$ at $\gamma = 0.3$, therefore, the null hypothesis H_0 concerning compatibility of this group of the PT results was not rejected with a probability of a type I error (to reject the hypothesis when it is correct) of 0.025 and a probability of a type II error (to not reject the hypothesis when it is false) of above 0.70 according to Fig. 8. Although not all individual z-score values were successful, the compatibility (group) assessment is positive, probably because of the standard [45] use.

EXAMPLE 5. SELECTION OF THE MOST COMMUTABLE (ADEQUATE) CRM FOR PT OF CEMENTS

B-5.1 Twelve components

In certificates of CRMs of Portland cements SRMs No. 1881, 1884–1888 (developed by NIST, USA) one can find the following caution: "to obtain the most accurate results by X-ray fluorescence methods of analysis, the user should compare his samples to the particular SRM that is most nearly the same in overall chemical composition". This caution means that all 12 certified components of the cement compositions (n = 12) should be taken into account. The expected composition of the sample under analysis and the certified values of the SRMs in % as mass fraction, as well as the ratio R_i and the adequacy score A_S values calculated as explained in paragraph 3.1.1 of the Guide, are shown in Table 10, where $c_{i,s}$ and $c_{i,cert}$ are the concentrations of the i-th component in a sample and in the CRM, respectively, and a_i is the sensitivity coefficient.

Table 10 Evaluation of commutability of SRMs No. 1881, 1884–1888 to a sample of cement.

i	Analyte	$c_{i,s}$	SRM 1881		SRM 1884		SRM 1885		SRM 1886		SRM 1887		SRM 1888	
			$\frac{100}{c_{i,cert}}$	R_i	$\frac{100}{c_{i,cert}}$	R_i	$\frac{100}{c_{i,cert}}$	R_i	$\frac{100}{c_{i,cert}}$	R_i	$\frac{100}{c_{i,cert}}$	R_i	$\frac{100}{c_{i,cert}}$	R_i
1	CaO	64	58.67	0.92	64.01	1.00	62.14	0.97	67.43	0.95	62.88	0.98	63.78	1.00
2	SiO_2	21	22.25	0.94	23.19	0.91	21.24	0.99	22.53	0.93	19.98	0.95	20.86	0.99
3	Al_2O_3	4	4.16	0.96	3.31	0.83	3.68	0.92	3.99	1.00	5.59	0.72	5.35	0.75
4	Fe_2O_3	4	4.68	0.85	3.30	0.83	4.40	0.91	0.31	0.08	2.16	0.54	3.18	0.80
5	SO_3	3	3.65	0.82	1.67	0.56	2.22	0.74	2.04	0.68	4.61	0.65	3.16	0.95
6	MgO	3	2.63	0.88	2.32	0.77	4.02	0.75	1.60	0.53	1.26	0.42	0.71	0.24
7	K_2O	0.5	1.17	0.43	0.51	0.98	0.83	0.60	0.16	0.32	1.27	0.39	0.56	0.89
8	TiO_2	0.2	0.25	0.80	0.16	0.80	0.20	1.00	0.19	0.95	0.27	0.74	0.29	0.69
9	Na_2O	0.1	0.04	0.40	0.13	0.77	0.38	0.26	0.02	0.20	0.10	1.00	0.14	0.71
10	SrO	0.1	0.11	0.91	0.048	0.48	0.037	0.37	0.11	0.91	0.07	0.70	0.07	0.70
11	P_2O_5	0.1	0.09	0.90	0.12	0.83	0.10	1.00	0.025	0.25	0.075	0.75	0.085	0.85
12	Mn_2O_3	0.1	0.26	0.38	0.11	0.91	0.12	0.83	0.013	0.13	0.072	0.72	0.025	0.25
A_S	$n = 12$			2.2		5.8		2.2		0.0		1.1		0.9
A_S	$n = 6$			51.2		26.6		44.3		2.5		9.9		13.2
A_S	$n = 2$			80.4		77.3		72.5		50.6		41.3		23.6
A_S	$a_1 = 0.9$			81.1		77.3		72.7		50.9		41.3		23.6

One can see that the SRM most commutable/adequate to the sample is SRM 1884 (A_S = 5.8 %), while the least suitable for the sample analysis is SRM 1886 (A_S = 0.004 %) at all a_i =1 [50]. The reason for such a low A_S value for SRM 1886 is the significant difference in concentrations of the components in the SRM and in the sample, especially for Fe_2O_3 and Mn_2O_3 (R_4 = 0.08 and R_{12} = 0.13).

B-5.2 Six components

When only the first 6 major components in Table 10 should be taken into account (n = 6), the SRM most adequate to the sample is SRM 1881 (A_S = 51.2 %), while the SRM least suitable for the sample analysis is again SRM 1886 (A_S = 2.5 %).

B-5.3 One component

At SRM selection for control of MgO gravimetric determination in the sample, only the CaO concentration in addition to the MgO content is important, since a small amount of CaO remains in the MgO precipitate. For this purpose, the most adequate SRM to the sample is SRM 1881 (A_S = 80.4 %) and the least suitable for the sample analysis is SRM 1888 (A_S = 23.6 %). The reason is that the content of MgO in SRM 1881 is the closest to the expected value in the sample under analysis, while in SRM 1888 it is the least close. The influence of the differences in CaO concentrations in the SRMs on their adequacy score is relatively less significant here.

B-5.4 Sensitivity coefficient

When the component influence is less significant, the sensitivity coefficient can be decreased. For example, the score of SRM 1881 adequacy to the sample at CaO sensitivity coefficient a_1 = 0.9 is A_S = 81.1 % (see the last line in Table 10). It is a little more than previously at a_1 = 1, when A_S = 80.4 % was obtained. The score of SRM 1888 (A_S = 23.6 %) is not changed here practically at all.

Naturally, choosing components and parameters for another analytical task will yield other score A_S values and, correspondingly, a different SRM selection.

MEMBERSHIP OF SPONSORING BODIES

Membership of the IUPAC Analytical Chemistry Division Committee for the period 2008–2009 was as follows:

***President*:** A. Fajgelj (Austria); ***Vice-President*:** W. Lund (Norway); ***Past-President*:** R. Lobinski (France); ***Secretary*:** D. B. Hibbert (Australia); ***Titular Members*:** M. F. Camões (Portugal); Z. Chai (China); P. De Bièvre (Belgium); J. Labuda (Slovakia); Z. Mester (Canada); S. Motomizu (Japan); ***Associate Members*:** P. De Zorzi (Italy); A. Felinger (Hungary); M. Jarosz (Poland); D. E. Knox (USA); P.Minkkinen (Finland); P.M. Pingarrón (Spain); ***National Representatives*:** S. K. Aggarwal (India); R. Apak (Turkey); M. S. Iqbal (Pakistan); H. Kim (Korea); T. A. Maryutina (Russia); R. M. Smith (UK); N. Trendafilova (Bulgaria).

Membership of the IUPAC Interdivisional Working Party on Harmonization of Quality Assurance was as follows:

***Chair*:** P. De Zorzi (Italy); ***Members*:** P. Bode (Netherlands); P. De Bièvre (Belgium); R. Dybkaer (Denmark); S. L. R. Ellison (UK); D. B. Hibbert (Australia); I. Kuselman (Israel); J. Y. Lee (Korea); L. Mabit (Austria); P. Minkkinen (Finland); U. Sansone (Austria); M. Thompson (UK); R. Wood (UK).

Membership of the Cooperation of International Traceability in Analytical Chemistry (CITAC) was as follows:

***Chairman*:** I. Kuselman (Israel); ***Vice Chairman*:** W. Louw (South Africa); ***Secretary*:** P. Charlet (France); ***Members*:** C. Puglisi (Argentina); A. Squirrell (Australia); L. Besley (Australia); A. Fajgelj (Austria); W. Wegscheider (Austria); P. De Bièvre (Belgium); H. Emons (Belgium); O. P. De Oliveira Junior (Brazil); V. Poncano (Brazil); G. Massiff (Chile); Y. Yadong (China); M. Suchanek (Czech Republic); I. Leito (Estonia); T. Hirvi (Finland); I. Papadakis (Greece); T. L. Ting (China); P. K. Gupta (India); M. Walsh (Ireland); K. Chiba (Japan); H. Y. So (Korea); Y. M. Nakanishi (Mexico); L. Samuel (New Zealand); V. Baranovskaya (Russia); Y. Karpov (Russia); C. Cherdchu (Thailand); R. Kaarls (Netherlands); S. L. R. Ellison (UK); M. Milton (UK); V. Iyengar (USA); C. Burns (USA); W. May (USA); J. D. Messman (USA); W. Wolf (USA); P. S. Unger (USA).

Membership of the Task Group was as follows:

***Chair*:** A. Fajgelj (Austria); ***Members*:** I. Kuselman (Israel); M. Belli (Italy); S. L. R. Ellison (UK); U. Sansone (Austria); W. Wegscheider (Austria).

ACKNOWLEDGMENTS

The Task Group would like to thank P. Fisicaro (France) and M. Koch (Germany) for their data used and help in preparation of Examples 1 and 2, respectively, in Annex B of the Guide; H. Emons (Belgium) and J. Lorimer (Canada) for helpful discussions; Springer, Heidelberg (<www.springer.com>) and the Royal Society of Chemistry, London (<www.rsc.org>) for permission to use material from the published papers cited in the Guide.

REFERENCES

1. M. Thompson, R. Wood. *Pure Appl. Chem.* **65**, 2123 (1993).
2. M. Thompson, S. L. R. Ellison, R. Wood. *Pure Appl. Chem.* **78**, 145 (2006).
3. ISO 13528. *Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons* (2005).

4. ISO/IEC Guide 43. *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* (1997).
5. ISO/IEC 17043. *Conformity Assessment: General Requirements for Proficiency Testing* (2010).
6. ILAC G13. *Guidelines for the Requirements for the Competence of Providers of Proficiency Testing Schemes* (2000).
7. A. Deom, R. E. L. Aouad, C. C. Heuck, S. Kumari, S. M. Lewis, A. Uldall, A. J. Wardle. *Requirements and Guidance for External Quality Assessment Schemes for Health Laboratories*, World Health Organization, WHO/DIL/LAB/99.2 (1999).
8. M. Belli, S. L. R. Ellison, A. Fajgelj, I. Kuselman, U. Sansone, W. Wegscheider. *Accred. Qual. Accur*. **12**, 391 (2007).
9. ISO/IEC 3534. *Statistics, Vocabulary and Symbols, Part 1: General Statistical Terms and Terms Used in Probability* (2006).
10. ISO/IEC Guide 99. *International Vocabulary of Metrology: Basic and General Concepts and Associated Terms* (VIM) 3rd ed. (2007). JCGM 200:2008 at <http://www.bipm.org/en/publications/guides/vim>.
11. EURACHEM/CITAC Guide. *Traceability in Chemical Measurement. A Guide to Achieving Comparable Results in Chemical Measurement* (2003).
12. ISO/IEC 17025. *General Requirements for the Competence of Testing and Calibration Laboratories* (2005).
13. P. Armishaw, B. King, R. G. Millar. *Accred. Qual. Assur*. **8**, 184 (2003).
14. W. Hasselbarth. In *Reference Materials in Analytical Chemistry. A Guide for Selection and Use*, pp. 16–18, A. Zschunke (Ed.), Springer, Berlin (2000).
15. D. B. Hibbert. *Accred. Qual. Accur.* **11**, 543 (2006).
16. I. Kuselman, M. Belli, S. L. R. Ellison, A. Fajgelj, U. Sansone, W. Wegscheider. *Accred. Qual. Accur*. **12**, 563 (2007).
17. H. Emons, A. Fajgelj, A. M. H. Van der Veen, R.Watters. *Accred. Qual. Assur.* **10**, 576 (2006).
18. V. I. Dvorkin. *Accred. Qual. Assur.* **9**, 421 (2004).
19. I. Kuselman, A. Weisman, W. Wegscheider. *Accred. Qual. Assur*. **7**, 122 (2002).
20. I. Ekeltchik, E. Kardash-Strochkova, I. Kuselman. *Microchim. Acta* **141**, 195 (2003).
21. A. Weisman, Y. Gafni, M. Vernik, I. Kuselman. *Accred. Qual. Assur*. **8**, 263 (2003).
22. I. Kuselman, M. Pavlichenko. *Accred. Qual. Assur*. **9**, 387 (2004).
23. W. J. Youden, E. H. Steiner. *Statistical Manual of the Association of Official Analytical Chemists*, AOAC, Arlington, VA, USA (1990).
24. D. Rizkov, O. Lev, J. Gun, B. Anisimov, I. Kuselman. *Accred. Qual. Assur*. **9**, 399 (2004).
25. P. De Bièvre. *Accred. Qual. Assur*. **11**, 487 (2006).
26. M. Sargent, G. Holcombe. *VAM Bull.* **34**, 19 (2006).
27. I. Kuselman. In *Combining and Reporting Analytical Data*, A. Fajgelj, M. Belli, U.Sansone (Eds.), pp. 229–239, RSC Special Publication No. 307, Royal Society of Chemistry, Cambridge (2006).
28. I. Kuselman. *Accred. Qual. Assur*. **10**, 466 (2006).
29. ASTM Standard D 2795-74. *Standard Methods of Analysis of Coal and Coke Ash* (1974).
30. I. Kuselman. *Accred. Qual. Assur*. **10**, 659 (2006).
31. D. B. Owen. *Handbook of Statistical Tables*, Addison Wesley, London (1962).
32. R. B. D'Agostino, M. A. Stephens (Eds.). *Goodness-Of-Fit Techniques*. Marcel Dekker, New York (1986).
33. Council Directive 1999/30/EC Relating to Limit Values for Sulfur Dioxide, Nitrogen Dioxide and Oxides of Nitrogen, Particulate Matter and Lead in Ambient Air.
34. Directive 2004/107/EC of the European Parliament and of the Council Relating to Arsenic, Cadmium, Mercury, Nickel and Polycyclic Aromatic Hydrocarbons in Ambient Air.

35. EN 14902. *Ambient Air Quality. Standard Method for the Measurement of Pb, Cd, AS, and Ni in the PM 10 Fraction of Suspended Particulate Matter* (2005).
36. EN 12341. *Air Quality: Determination of the PM10 Fraction of Suspended Particulate Matter: Reference Method and Field Test Procedure to Demonstrate Reference Equivalence of Measurement Methods* (1998).
37. ISO 5725-5. *Accuracy (Trueness and Precision) of Measurement Methods and Results, Part 5: Alternative Methods for the Determination of the Precision of a Standard Measurement Method* (1998).
38. O. Rienitz. *Entwicklung Chemisch-Analytischer Primärmethoden zur Bestimmung Physiologisch Relevanter Anorganischer Bestandteile in Humanserum,* Dissertation, PTB-Bericht PTB-ThEx-19, Braunschweig (2001).
39. L. Kimhi, C. Zlotnikov, I. Kuselman. *Accred. Qual. Assur*. **11**, 577 (2006).
40. Israeli Standard 26. *Part 4: Methods of Testing Concrete: Strength of Hardened Concrete* (1985).
41. Israeli Standard 118. *Concrete for Structural Uses: Production Control and Compressive Strength* (2003).
42. ASTM Standard D 3339. *Standard Test Method for Acid Number of Petroleum Products by Semi-Micro Color Indicator Titration* (2007).
43. ISO 6618. *Petroleum Products and Lubricants: Neutralization Number: Colour-indicator Titration Method* (1997).
44. ISO 6619. *Petroleum Products and Lubricants: Neutralization Number: Potentiometric Titration Method* (1988).
45. ASTM Standard D 664. *Standard Test Method for Acid Number of Petroleum Products by Potentiometric Titration* (2007).
46. I. Kuselman, Ya. I. Tur'yan, E. Strochkova, I. Goldfeld, A. Shenhar. *J. AOAC Int.* **83**, 282 (2000).
47. E. Kardash-Strochkova, Ya. I. Tur'yan, I. Kuselman, N. Brodsky. *Accred. Qual. Assur*. **7**, 250 (2002).
48. ASTM D445. *Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (the Calculation of Dynamic Viscosity)* (2006).
49. I. Kuselman. *Accred. Qual. Assur*. **5**, 100 (2000).
50. I. Kuselman. *Accred. Qual. Accur*. **9**, 591 (2004).

ABBREVIATIONS AND SYMBOLS

A	critical value for numbers N_+ and/or N_-
AAS	atomic absorption spectrometry
a_i	empirical sensitivity coefficient of the i-th component
A_N	acid number
A_S	adequacy score
b_{cf}	buoyancy correction factor
c_1, c_2	measurement/test results corresponding to the crossing points of two probability density functions
c_{cert}	certified (assigned) value of a particular property of a CRM
c_i	measurement/test result of i-th laboratory participating in PT
c_{is}	value of a particular property of routine samples
c_{PT}	population (theoretical) mean of PT results
$c_{PT/avg}$	observed/experimental mean of PT results (consensus value)
CRM	certified reference material
EMD	Ecole des Mines de Douai, France
E_n, z, and ζ	scores for assessment of proficiency of a laboratory participating in PT
F	frequency of a c-value

f	probability density function
GC-MS	gas chromatography-mass spectroscopy
GF-AAS	graphite furnace-atomic absorption spectrometry
H_0	null hypothesis
H_1	alternative hypothesis
hand	hand preparation of a sample
HPLC	high-performance liquid chromatography
i, j, n	index numbers
ICP-MS	inductively coupled plasma mass spectroscopy
ICP-OES	inductively coupled plasma-optical emission spectroscopy
ID-ICP-MS	isotope dilution-inductively coupled plasma-mass spectrometry
IHRM	in-house reference material
INPL	National Physical Laboratory of Israel
ISO	International Organization for Standardization
K	kelvin
LNE	Laboratoire National de Métrologie et d'Essais, France
MCL	maximum contaminant level
$m_{As_2O_3}$	mass of a sample of arsenic oxide
m_{dil}	mass of the diluted solution (a sample)
$m_{dil/t}$	total mass of the diluted solution
m_{lot}	total mass of final lot
M_{PT}	population median of PT results
m_{ss}	mass of the stock solution (a sample)
$m_{ss/t}$	total mass of the stock solution
N	size of the statistical sample of measurement results of PT participants
N_-	number of PT results $c_i < c_{cert} - \Delta$
N^*	number of potentiometric titration results
N_+	number of PT results $c_i > c_{cert} + \Delta$
NIST SRM	standard (certified) reference material developed by the National Institute of Standards and Technology, USA
NMR	nuclear magnetic resonance
N_p	size of the population of PT participants
P	probability
$p_{As/As_2O_3} = M(As)/M(As_2O_3)$	proportion of the atomic or molecular weights M
P_C	power of criterion
p_c	purity of chemicals
P_e	probability of an event
P_T	power of test
pH-metr.	pH-metric method
Pot. titr.	potentiometric titration
PT	proficiency testing
Quest	questionable
R_{A_N}	limit of a difference between two results of A_N determination (range)
R_i	ratio of the min to the max values from two concentrations
RL	reference laboratory
s	observed sample standard deviation
SADCMET	Southern African Development Community Cooperation in Measurement Traceability
s_{bsi} and s_{isi}	between- and intra-sample standard deviations
SI	International System of Units

s_{PT}	observed sample standard deviation of PT results
t	quantile of the Student's distribution
$u(c_i)$ and $U(c_i)$	standard and expanded uncertainties of c_i, respectively
u_{cert} and U_{cert}	standard and expanded uncertainty of c_{cert}, respectively
u_{comb}	combined standard uncertainty
u_{mLP}	standard measurement uncertainty declared by a laboratory participating in PT
u_{mRL}	standard measurement uncertainty declared by the reference laboratory
USN	ultrasonic nebulization
UV	ultraviolet
vibr	sample preparation with a vibrating table
VIM3	*International Vocabulary of Metrology*, 3rd ed.
x_j	normalized value of the j-th PT result
z	see E_n
α	significance level; probability of a type I error
β	probability of a type II error
γ	ratio $\sigma_{cert}/\sigma_{PT}$
Δ	permissible bias of M_{PT} from c_{cert}
δ and λ	parameters
Π	product
ρ_{lot}	density of a lot of an aqueous IHRM
σ_{PT}	population standard deviation of PT results
$\sigma_{PT/av}$	standard deviation of the sample mean $c_{PT/av}$ of PT results
σ_{targ}	target standard deviation of PT results
χ^2	quantile of the χ^2 distribution
ϕ	function of normalized normal distribution function
$\phi(x_j)$	value of the function of the normalized normal distribution for x_j
φ	fraction of the statistical sample of size N from the population of size N_p
ω^2	empirical value of the Cramer-von-Mises criterion
ζ	see E_n