Eurachem

A focus for analytical chemistry in Europe

Selection, Use and Interpretation of Proficiency Testing (PT) Schemes

Third Edition 2021

Eurachem

Selection, Use and Interpretation of Proficiency Testing (PT) Schemes

Third Edition 2021

Editors

Brian Brookman (LGC, UK)
Ian Mann (SAS, CH)

Composition of Working Group*

Brian Brookman (Chair), LGC, UK
Natalie Mason (Secretariat), LGC, UK
Stéphanie Albarède, CTCB, FR
Frank Baumeister, ISWA, DE
Ute Braun, MUVA, DE
Laura Ciaralli, ISS, IT
Dana Dominova, CAI, CZ
Magnus Holmgren, RISE, SE
Hans D. Jensen, DANAK, DK
Christian Lehmann, DAkkS, DE
Ulrich Leist, DRRR, DE
Mirja Leivuori, SYKE, FI
Caroline Leonard, Merck, FR
Ian Mann, SAS, CH
Richard McFarlane, UKAS, UK
Raquel Murtula Corbi, LABAQUA, ES
Sabrina Pepa, Accredia, IT
Erika Sarkany, QualiCont, HU
Anne Vegard Stavelin, Noklus, NO
Johannes van de Kreeke, BAM, DE
**At time of document approval*

Acknowledgements

This edition has been produced by the EEE-PT Working Group (EA-Eurolab-Eurachem). EQALM (European Organisation for External Quality Assurance Programmes in Laboratory Medicine) as an affiliate member was also involved in this revision.

Citation

This publication should be cited* as:
"B. Brookman and I. Mann (eds.) Eurachem Guide: Selection, Use and Interpretation of Proficiency Testing (PT) Schemes (3rd ed. 2021). Available from www.eurachem.org."

**Subject to journal requirements*

Selection, Use and Interpretation of Proficiency Testing (PT) Schemes

English edition

3rd edition 2021

Contents

Abbreviations and symbols

The following abbreviations and symbols occur in this guide.

Abbreviations

CIPM	International Committee for Weights and Measures
CRM	certified reference material
DIN	Deutsches Institut für Normung (German Institute of Standardization)
EA	European Co-operation for Accreditation
EPTIS	international database of proficiency testing schemes
EQA	External quality assessment
EU	European Union
ICP-MS	inductively coupled plasma mass spectrometry
IEC	International Electrotechnical Commission
IFCC	International Federation of Clinical Chemistry and Laboratory Medicine
ILAC	International Laboratory Accreditation Cooperation
ILC	interlaboratory comparison
IQC	internal quality control
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
MRA	mutual recognition arrangement
PCR	polymerase chain reaction
PT	proficiency testing
QA	quality assurance
QC	quality control
RM	reference material
SDPA	standard deviation for proficiency assessment

Symbols

E_n	E_n performance score
s^*	robust estimate of the participant standard deviation
σ_{pt}	SDPA
p	number of participants
x_i	result reported by participant i
x_{pt}	assigned value
$u(x_i)$	standard uncertainty of a result from participant i
$u(x_{pt})$	standard uncertainty of the assigned value
$U(x_i)$	expanded uncertainty of reported result from participant i
$U(x_{pt})$	expanded uncertainty of the assigned value

z	z performance score
z'	modified z (prime) performance score
ζ	zeta performance score

1 Introduction

A regular independent assessment of the technical performance of a laboratory is necessary to assure the validity of results, and should be part of an overall quality strategy. A common approach to this independent assessment is the use of independent proficiency testing (PT) schemes or external quality assessment (EQA) schemes as they are often referred to within the medical sector. A PT scheme is a system for objectively evaluating a laboratory's performance by the use of external means, and includes regular comparison of a laboratory's measurement results with those of other laboratories. This is achieved by the PT provider distributing sufficiently homogeneous and stable PT items to participants for analysis and reporting of the measurement results. Each distribution of PT items is referred to as a "round". The main objective of a PT scheme is to help the participant to assess the validity of its measurements. PT schemes may also address pre-analytical and post-analytical aspects of the measurement cycle, i.e. specific procedures carried out before or after the analytical procedure. These may, for example, include sampling or the interpretation of the measurement results. In addition, participation in an appropriate PT scheme, including participation in small interlaboratory comparisons (ILCs) [1] where an appropriate PT scheme is not available, is required for laboratories seeking recognition of their competence through accreditation against the standards ISO/IEC 17025 [2] or ISO 15189 [3]. In some sectors participation in specific schemes can be mandatory. PT schemes are operated for the benefit of participants. However, other parties also have a legitimate interest in PT schemes. These include customers of analytical laboratory services, accreditation bodies, regulatory authorities and other end-users of laboratory results. It is important for PT providers to bear in mind the needs of these organizations in order that they are able to use the results from PT schemes to aid their understanding of the capabilities and competence of laboratories (See Annex C).

It is important for laboratories to have comprehensive information on the availability and scope of PT schemes in the areas in which they work. This will enable them to make appropriate decisions about which PT scheme(s) they should participate in. It is important that this type of information is widely available in order for laboratories to be able to select the most appropriate PT scheme.

Laboratories also need to have a good understanding of PT, what the objectives of the PT schemes are, any limitations, how the data is evaluated by the PT provider, and how the data from PT schemes should be internally evaluated and used.

In this Guide, the term "measurement" is used generally for measurements undertaken in both calibration and testing. The term "examination" is understood to include examination of both quantitative and qualitative characteristics (see ISO 15189). For brevity, unless otherwise qualified, provisions on measurement of a PT item, measurement procedures etc. should be read as applying equally to any laboratory activity leading to a reported proficiency test result, whether quantitative, qualitative or interpretive.

There are a number of key principles, covered in this document, which help to ensure the appropriateness of participation in PT schemes, and need to be considered and understood by interested parties:

a) the PT scheme in which a laboratory participates should resemble as closely as possible the laboratory's routine work, for example, in terms of sample matrix, characteristics and levels; any differences should be noted and accounted for;

b) laboratories should treat PT items as routine samples;

c) the evaluation and interpretation of the performance in a PT scheme should take into account the risk associated with the measurement;

d) all unsatisfactory or repeated questionable results must be thoroughly investigated so that the laboratory can understand the reasons for poor performance and correct as necessary;

e) the performance of a laboratory over several rounds of a PT scheme and analysis of trends is paramount to determining the successfulness of participation;

f) the PT scheme documentation, such as scheme protocols, must provide clear information in order for all parties to understand how the PT scheme operates;

g) the PT provider should be open to discussion amongst interested parties in order to gain a more accurate understanding of the PT scheme and its operation;

h) laboratories should view PT participation as an educational tool, using the PT scheme results in the improvement process and to give feedback to staff.

2 Scope

The aim of this document is to give laboratories guidance on:

a) aims and benefits of participation in PT schemes;

b) selecting the most appropriate PT scheme;

c) understanding the basic statistics and performance scoring used by the PT providers;

d) using and interpreting the PT results in order to improve the overall performance of the laboratory.

This document is aimed at all organizations that are performing sampling, testing, calibrations and examinations, for example testing laboratories, calibration laboratories, inspection bodies, biobanks, etc. It covers measurements, examinations and interpretations.

This document does not address those ILCs that are aimed at the evaluation of performance characteristics of a measurement procedure (e.g. “ring trials” or “collaborative studies”) or the assignment of values to reference materials (e.g. “certification” of RMs), or comparisons conducted in the framework of the CIPM MRA [4], such as the “key comparisons” that are aimed at National Metrology Institutes.

3 Definitions

3.1 Interlaboratory comparison (ILC)

Organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions

[ISO/IEC 17043, definition 3.4] [5]

3.2 Proficiency testing (PT)

Evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons

NOTE Some providers of proficiency testing in the medical area use the term External Quality Assessment for their proficiency testing schemes and/or for their broader programmes.

[ISO/IEC 17043, definition 3.7] [5]

3.3 Proficiency testing scheme (PT scheme)

Proficiency testing designed and operated in one or more rounds for a specified area of testing, measurement, calibration or inspection

[ISO/IEC 17043, definition 3.11] [5]

3.4 Proficiency test item (PT item)

Sample, product, artefact, reference material, piece of equipment, measurement standard, data set or other information used for proficiency testing

[ISO/IEC 17043, definition 3.8] [5]

3.5 Proficiency testing provider (PT provider)

Organization which takes responsibility for all tasks in the development and operation of a proficiency testing scheme

[ISO/IEC 17043, definition 3.9] [5]

3.6 Participant

Laboratory, organization or individual, that receives proficiency test items and submits results for review by the proficiency testing provider

NOTE In some cases the participant can be an inspection body.

[ISO/IEC 17043, definition 3.6] [5]

3.7 Assigned value

Value attributed to a particular property of a proficiency test item

[ISO/IEC 17043, definition 3.1] [5]

NOTE for pre- and post- analytical PT schemes this could for example take the form of an expert opinion – see 3.13

3.8 Standard deviation for proficiency assessment (SDPA)

Measure of dispersion used in the evaluation of results of proficiency testing, based on the available information

NOTE 1 The standard deviation applies only to ratio and differential scale results.

NOTE 2 Not all proficiency testing schemes evaluate proficiency based on the dispersion of results.

[ISO/IEC 17043, definition 3.13] [5]

3.9 Measurement

Process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity

NOTE 1 Measurement does not apply to nominal properties.

NOTE 2 Measurement implies comparison of quantities or counting of entities.

NOTE 3 Measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions.

[JCGM 200:2012, definition 2.1] [6]

3.10 Measurement uncertainty (uncertainty of measurement/uncertainty)

Non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used

NOTE 1 Measurement uncertainty includes components arising from systematic effects, such as components associated with corrections and the assigned quantity values of measurement standards, as well as the definitional uncertainty. Sometimes estimated systematic effects are not corrected for but, instead, associated measurement uncertainty components are incorporated.

NOTE 2 The parameter may be, for example, a standard deviation called standard measurement uncertainty (or a specified multiple of it), or the half-width of an interval, having a stated coverage probability.

NOTE 3 Measurement uncertainty comprises, in general, many components. Some of these may be evaluated by Type A evaluation of measurement uncertainty from the statistical distribution of the quantity values from series of measurements and can be characterized by standard deviations. The other components, which may be evaluated by Type B evaluation of measurement uncertainty, can also be characterized by standard deviations, evaluated from probability density functions based on experience or other information.

NOTE 4 In general, for a given set of information, it is understood that the measurement uncertainty is associated with a stated quantity value attributed to the measurand. A modification of this value results in a modification of the associated uncertainty.

[JCGM 200:2012, definition 2.26] [6]

3.11 Measurement procedure

Detailed description of a measurement according to one or more measurement principles and to a given measurement method, based on a measurement model and including any calculation to obtain a measurement result.

NOTE 1 A measurement procedure is usually documented in sufficient detail to enable an operator to perform a measurement.

NOTE 2 A measurement procedure can include a statement concerning a target measurement uncertainty.

NOTE 3 A measurement procedure is sometimes called a standard operating procedure, abbreviated SOP.

[JCGM 200:2012, definition 2.6] [6]

NOTE 4 EA-4/18 [7] uses the equivalent term “Measurement process”, to indicate: “The process of measuring the characteristic, including any pre-treatment required to present the sample, as received by the laboratory, to the measuring device”.

3.12 Examination

Set of operations having the object of determining the value or characteristics of a property

[ISO 15189] [3]

3.13 Characteristic

The parameter being measured

[EA-4/18] [7]

NOTE 1 Examples: arsenic, fat, creatinine, length, hardness, force e

NOTE 2 In the context of this document, this can be further expanded to include other characteristics, such as opinions, colour, taste, presence/absence.

3.14 Measurand

Quantity intended to be measured

[JCGM 200:2012, definition 2.3, excluding notes] [6]

3.15 Product

The item to which the measurement process is being applied

[EA-4/18] [7]

NOTE Examples: soil, vegetables, serum, polystyrene, concrete

3.16 Area of technical competence

Field of expertise defined by a minimum of one measurement process, characteristic and product, which are related Example: amount of arsenic in soil by ICP-MS

[EA-4/18] [7]

3.17 Level of participation

The number of specific activities that an organisation identifies within its scope of accreditation, and therefore the number of specific proficiency tests that should be considered for participation

[EA-4/18] [7]

3.18 Frequency of participation

The number of proficiency tests per unit of time, in which a laboratory participates for an activity as specified in their scope of accreditation

[EA-4/18] [7]

4 Introduction to proficiency testing

4.1 Role of PT within the management system

In order to monitor the validity of its measurements, it is important for the laboratory to implement a quality assurance (QA) system, which includes the monitoring of its performance by comparison with results of other laboratories, where available and appropriate. For laboratories that are accredited, or seeking accreditation, these measures are an important aspect of the requirements. It is a requirement of ISO/IEC 17025 [2] and ISO 15189 [3] for laboratories to participate in PT or another ILC. In ISO 15189 only, the use of alternative approaches when an ILC is not available is specified. PT providers that meet the requirements of ISO/IEC 17043 [5] are considered to be competent.

PT plays a highly valuable role as it provides objective evidence in the monitoring of the competence of the participant. This evidence can be used to improve the performance of the participant and/or give confidence in the participant's ability to perform a specific measurement.

Furthermore, participation in PT schemes not only gives information on the performance of the measurement procedure, but also on other aspects of the management system such as sampling, reception/treatment of the sample, treatment of the data, interpretation of the results, result reporting, etc. It is most important that the laboratory sets up a relevant strategy for participation in PT schemes (see 5.2).

PT provides an opportunity to undertake comparisons of the participants' data with assigned values (or other performance criteria), or with the performance of peer laboratories. The results from a PT scheme will provide participants with either a confirmation that their performance is satisfactory or an alert that investigation of potential problems is required.

It is important to underline that the aim of PT participation is not just about the performance score, but also about enabling the participant to learn from their participation in PT schemes and to use this information to improve the quality of their measurements.

Although the main aim of a PT scheme is to evaluate the performance of participants, there are many other benefits, which are detailed in chapter 6.

4.2 Types of PT schemes

Various types of PT schemes are available, each based on at least one element of each of the following four features:

4.2.1 Type of expected result

a) qualitative: the results of qualitative tests are identified on categorical (nominal) or ordinal scales. This includes:

- PT schemes that require reporting on a categorical scale (sometimes called 'nominal'), where the characteristic value has no magnitude (such as a type of substance or organism);
- PT schemes for presence or absence of a characteristic, whether determined by subjective criteria or by the magnitude of a signal from a measurement procedure. This can be regarded as a special case of a categorical or ordinal scale, with only two values (also called 'dichotomous', or binary);
- PT schemes requiring results reported on an ordinal scale, which can be ordered according to magnitude but for which no arithmetic relationships exist among different results. For example, 'high, medium and low' form an ordinal scale.

b) quantitative: the results are numeric and are generally reported on an interval or a ratio scale;

NOTE 1 An interval scale is a measurement scale in which a certain distance along the scale means the same thing no matter where on the scale you are, but where "0" on the scale does not represent the absence of the parameter being measured and on which ratios do not have a consistent interpretation. Fahrenheit and Celsius temperature scales are examples of interval scales.

NOTE 2 A ratio scale is a measurement scale in which a certain distance along the scale means the same thing no matter where on the scale you are, and where "0" on the scale represents the absence of the parameter being measured, and on which ratios have a consistent meaning, e.g. a "4" on such a scale implies twice as much

of the parameter being measured as a "2" and “8” implies twice as much as “4”. The kelvin scale for temperature is an example of a ratio scale.

c) interpretive: no measurement is involved. The PT item is an object, a measurement result, a set of data or other set of information requiring judgement within the participant’s competence.

4.2.2 Frequency

a) single occasion exercise: PT items are provided as a one-off exercise;

b) continuous: PT items are provided on a regular basis.

4.2.3 Distribution format

1) sequential: the PT item to be measured is circulated successively from one participant to the next. In this case the PT item may be returned to the PT provider before being passed on to the next participant in order to determine whether any changes have taken place to the PT item. It is also possible for the participants to converge in a common location to measure the same PT item;

2) simultaneous: in the most common PTs, randomly selected sub-samples from a homogeneous bulk material are distributed simultaneously to participants for concurrent measurement. After reception of the results the PT provider will evaluate, on the basis of statistical techniques, the performance of each individual participant and of the group as a whole.

4.2.4 Processes

a) pre-analytical [8]: in this type of PT scheme, the PT item can be an object (e.g. a toy), on which the participant has to decide which measurements should be conducted, or a set of data or other information (e.g. a case study). This may also be the sampling procedure that is used to collect the samples for measurement within the laboratory;

b) analytical: the focus is specifically on the analytical process;

c) post-analytical: in this type of PT scheme, the PT item can be a set of data on which the participant is requested to give an opinion or interpretation.

These three processes represent the three phases in a complete measurement cycle; a PT scheme could be designed to address one or more of the phases or parts of a particular phase.

One special application of PT, often called “blind” PT, is where the PT item is indistinguishable from normal customer items or samples received by the participant. All types of PT schemes mentioned above could be organised as a blind PT.

5 Selection of appropriate PT schemes

5.1 Introduction

The selection of a PT scheme from a competent PT provider is critical to ensure that the participant obtains the most benefit from taking part in the PT scheme. It is therefore essential that the participant evaluates the competence of the PT provider.

Participating in a PT scheme supplements a laboratory's own internal quality control (IQC) procedures by providing an additional external measure of its measuring capability. Thus, all laboratories need to establish an adequate PT participation strategy with the aim of participating in relevant PT schemes, at a frequency appropriate to their circumstances [7].

In selecting the appropriate PT scheme, within an area of technical competence, a laboratory should answer at least the following questions to ensure that the PT scheme is fit for purpose (see 5.4):

1) What level of PT and frequency do I need?
2) Do any PT schemes exist for the various areas of technical competence?
3) Is the PT scheme relevant?
4) Is the PT provider competent, e.g. does the PT provider operate according to ISO/IEC 17043 [5]?
5) Is the selected PT scheme independent of any manufacturing or marketing interests in equipment, reagents or calibrators in its field of operation?

In cases where participation in PT is a mandatory requirement specified by regulatory authorities, the laboratory may have no choice regarding PT scheme selection and frequency of participation, although a choice of PT provider may be available.

5.2 Strategy of PT participation

Before selecting a PT scheme, laboratories should evaluate the level and frequency of their participation and establish their PT participation strategy. This evaluation should be done by taking into account the various areas of technical competence of the laboratory. The laboratory can then select the most appropriate PT scheme from a competent PT provider. A PT provider that operates according to ISO/IEC 17043 can be considered as competent, but a laboratory would need to verify this, through appropriate means (e.g. on-site witnessing, remote audit, document review etc.), if the PT provider is not accredited. Accreditation of PT providers against ISO/IEC 17043 provides the necessary evidence to a laboratory.

The laboratory should define its level and frequency of participation after careful analysis of its other QA measures (especially those that are able to disclose, quantify and follow the development of bias of a stated magnitude). The level and frequency of participation should depend on the extent to which other measures have been taken. Other types of QA measures include, but are not limited to:

- regular use of (certified) reference materials ((C)RMs);
- comparison of analysis by independent measurement procedures;
- participation in method development/validation and/or RM characterisation studies;
- use of IQC measures;
- other interlaboratory or intralaboratory comparisons, e.g. analysis of blind samples within the laboratory.

It must be recognised that there are sectors where participation in PT may be difficult, due to the technical characteristics of the measurement, the lack of PT schemes, the low number of existing laboratories in the sector, etc. Sometimes PT may only be possible or economically feasible for parts of the measurement procedure undertaken. In these cases the suitability of other QA/QC measures is paramount. Any legislative requirements for level and frequency of participation should be considered by the laboratory.

Laboratories should be able to justify and, where required, document the technical arguments that have led to their decision on the level and frequency of participation in PT.

5.2.1 Level of participation

The laboratory should first establish its areas of technical competence. The areas of technical competence of a laboratory can be defined by one measurement procedure, one characteristic and one product. Some areas may contain more than one measurement procedure, characteristic or product as long as equivalence and comparability can be demonstrated.

When determining an area of technical competence, it may be helpful to consider a stepwise approach working up from measurement procedure through characteristic to products. This is because it is more likely that there will be several products and/or characteristics associated with one measurement procedure within a given area of technical competence than vice versa. With reference to the:

a) measurement procedure: it is possible but not common to include different measurement procedures, as long as they can be considered as equivalent, in the same area of technical competence;

b) characteristic, that is the parameter to be measured, determined or identified: it may be possible to include more than one characteristic in the same area of technical competence;

c) products to be analysed: it may be possible to include different products in the same area of technical competence provided that the matrices, objects or materials included, are of equivalent nature.

Once the laboratory has defined its areas of technical competence the “level of participation” can be deemed to have been defined.

5.2.2 Frequency of participation

The laboratory should consider the level of risk affecting the laboratory, the sector in which it operates or the measurement procedures it is using. This can be determined, for example, by considering:

- number of measurements undertaken;
- turnover of technical staff;
- experience and knowledge of technical staff;
- source of metrological traceability (e.g. CRMs, national standards);
- known instability of the measurement procedure;
- significance and final use of measurement data (e.g. forensic science represents an area requiring a high level of assurance).

NOTE EA 4/18 [7] provides guidance and some case studies illustrating the choice of levels and frequency of participation

5.3 Availability of PT schemes

Information on PT providers and/or the availability of PT schemes can be found by various means:

a) Various international databases that list available PT schemes, for example:
   - EPTIS database [9] which lists hundreds of PT schemes operated around the world;
   - IFCC database [10] which lists many PT schemes in the medical sector;

b) National accreditation bodies can provide details of accredited PT providers and their associated scope;

c) Peer laboratories that already participate in, or know about relevant PT schemes;

d) The PT providers in the laboratory’s own country may also have information about the PT schemes of other providers;

e) A search on the Internet, using relevant keywords, can also provide useful information.

5.4 How to decide if the selected PT scheme is fit for purpose

If similar PT schemes are available and a choice has to be made, the laboratory should take into account that different PT schemes will provide different degrees of fitness for purpose, and that it is rare that a PT scheme that is a perfect fit exists. Therefore, in practice, the PT scheme that provides the best fitness for purpose has to be chosen.

The criteria used to determine the fitness for purpose of the PT scheme should include, but are not limited to, the following questions:

a) Are the number and size of PT items provided by the PT provider appropriate for the tests being carried out?

b) Are the types of PT items and/or levels/concentrations offered by the PT scheme similar to those encountered in the laboratory?

c) Is the statistical design described and does it take different measurement procedures into consideration?

NOTE 1 Statistical design covers the process of planning, collection, performance evaluation and reporting of the PT scheme data.

NOTE 2 Many PT providers have examples of their PT reports and/or copies of the scheme protocol on their websites, which makes it possible to review the performance evaluation (scoring) that is used.

NOTE 3 Are performance specifications adapted to the activity of the customer? Most PT schemes use performance specifications based on the state of the art (statistical dispersion) but in the medical field it is important to take into consideration the impact of the result on patient care. In this case performance specifications based on clinical outcome or biological variation (fixed acceptable limits) are preferable [11].

NOTE 4 Is the standard deviation for proficiency assessment (SDPA) being used by the PT provider fit for purpose for the laboratory?

d) Is the number and origin of participants for the PT scheme appropriate?

NOTE In some cases the results are dependent on the measurement procedure and the number of participants using a particular measurement procedure (peer group) has to be taken into account.

e) Is the frequency of rounds sufficient?

f) Does the PT provider have appropriate experience and are they competent?

NOTE Accreditation of a PT provider to ISO/IEC 17043 by an accreditation body provides evidence of their competence.

PT schemes have an important educational role. If the PT provider provides information on the performance of different measurement procedures and also takes into account the uncertainty in results reported by the participants, this can give valuable information for the further evaluation of measurement procedures.

It is important to note that it is the responsibility of the laboratory itself to decide about the criteria to be addressed, to make the comparison and to judge the relevancy of the PT scheme. The diagram in Annex A provides a helpful guide to performing this selection process. When assessing the PT scheme based on the questions included in the diagram, the laboratory should consider, for example, which aspects are mandatory, useful or not necessary. This should help to define if the PT scheme and the needs of the laboratory are sufficiently comparable. If they are, the laboratory should give serious consideration to participating in the PT scheme. A number of PT providers allow participation in just one round. If the laboratory is not fully convinced of the relevancy of the PT scheme this is a good option. Sometimes PT items from previous rounds of a PT scheme can be purchased together with the PT report. This is also a good option for judging the relevance of a specific PT scheme before electing to participate in all the rounds of the PT scheme.

6 Use of PT by laboratories

6.1 Introduction

In the context of this guide, the word laboratory implies all organizations performing testing or calibration activities, for example testing, calibration and medical laboratories, inspection bodies, biobanks and RM producers. The basic use of PT for a laboratory is to assess its performance for the conduct of specific tests, measurements or calibrations. Participating in a PT scheme provides the laboratory with the opportunity to compare their results with other laboratories through an independent external assessment.

The results and information received from the participation in PT schemes will provide laboratories with either a confirmation that their performance is satisfactory, or an indication that there are potential problems and that corrections should be made. To maximise the benefits of participating in PT schemes, it is essential that participants pay attention to the documents provided by the PT provider.

However, the use of PT should be much wider than the basic statement of whether the laboratory is competent or not. Laboratories can, as mentioned in 4.1, benefit from the participation in PT schemes in various ways. Some of these are listed below [12].

6.2 Identifying measurement problems

If a laboratory's result in a PT scheme indicates unsatisfactory performance, this should start a process of investigation of potential sources of error (see 8.3). Without participation in the PT scheme, such sources of error could remain undetected and the laboratory would not have been able to undertake appropriate corrective actions. This, in turn, could have resulted in the laboratory continuing to provide poor results to its customers or other stakeholders. Eventually, such errors could also lead to the loss of reputation of the laboratory or to legal or other action being taken by the customers or other stakeholders, such as regulatory bodies. In this regard the use of PT may be considered to be a risk management and quality improvement tool.

6.3 Comparing measurement procedures

For some laboratories, their participation in a PT scheme might be used to trial their performance using a new, modified or infrequently conducted measurement. In other cases, the participation may provide an opportunity to compare the results achieved by the laboratory using different measurement procedures (or when determining different concentration levels, etc.) to those normally used by the laboratory.

The PT report might, in some cases, provide summaries and comparisons of all the measurement procedures or commercial kits used by the participants, as is the case with the EU vigilance procedure related to in-vitro diagnostic medical devices [13]. For new or unusual activities, such data could be most valuable and assist the future selection of appropriate measurement procedures by the laboratory or indicate the need for additional investigation before adoption of new measurement procedures.

6.4 Comparing operator capabilities

If sufficient PT items are available to allow more than one operator within a laboratory to carry out the analysis, the laboratory has the added benefit of being able to compare the results of its operators. In addition, this might also provide some inputs to the laboratory's evaluation of its measurement uncertainty for the relevant measurements.

This might also allow the laboratory to compare the between-operator precision with published (or otherwise available) data for the measurements concerned.

The PT scheme itself might, in some cases, enable separate results of one or more operator to be reported.

6.5 Comparing analytical systems

PT results can provide an objective external assessment of the relative performance of analytical systems (on the same or different sites) used within a laboratory.

6.6 Improving performance

When a laboratory is not satisfied with its own results in a PT scheme, this provides an opportunity for the laboratory's management to investigate areas where its future testing could be improved. This might, for example, include additional operator training, adoption of new or modified measurement procedures, enhancing IQC of data, or equipment modifications, calibration or replacement (See section 8.3).

6.7 Educating staff

Many PT schemes have, as one of their objectives, provision of information on measurement procedures, data interpretation, uncertainty assignments, etc., which arise from the overall results in the PT scheme, or which are provided by experts involved in evaluating such results. Some PT schemes have a comprehensive educational role for participants and individual operators.

6.8 Exchange of information with the PT provider

Following the issue of a PT report, the laboratories usually have the possibility to contact the PT provider in order to obtain additional information about the results, or advice concerning the potential cause of non-satisfactory results. Some PT providers also hold "Participants' meetings", which can provide very useful information for the laboratories.

6.9 Instilling confidence in staff, management, and external users of laboratory services

Successful performance in a PT scheme can provide individual staff and their direct managers with additional confidence. Other management, including those without relevant technical expertise, can also be re-assured by the successful performance of their staff, often in areas of critical significance to their organization's activities and responsibilities.

External users of laboratory services, including their customers and the parties affected by the outcome of measurement, can also be given added confidence when made aware that a laboratory is willing to have its performance regularly evaluated through PT participation.

6.10 Measurement uncertainty

The laboratory's results from its participation in PT can, with caution, be used to check the evaluated measurement uncertainty, since that uncertainty should be compatible with the spread of results obtained by that laboratory over a number of PT rounds.

The "PT approach" can, in specific cases, also be used to evaluate the uncertainty. For example, if the same measurement procedure is used by all the participants in the PT scheme, the standard deviation is equivalent to an estimate of the reproducibility and can, in principle, be used in the same way as the reproducibility standard deviation obtained from an ILC undertaken for the purpose of performance characterisation of the measurement procedure [14 - 17]. Results from PT for sampling can be used to evaluate measurement uncertainty arising from between-sampler bias [18].

6.11 Use of PT items as Internal Quality Control materials

In some PT schemes, where there is sufficient, stable material provided to participants, the unused material could be used as a QC material for monitoring measurement performance as part of the laboratory's IQC procedures.

Where appropriate, the assigned values for the PT item might be considered useful as internal reference values for QC of measurement, operator training, etc.

A Eurachem information leaflet [19] gives additional information on using surplus PT materials.

6.12 Determining measurement precision and/or trueness

Depending on the design, some PT schemes will be useful in determining the precision (repeatability and reproducibility) or comparative trueness of the measurement procedures used in the PT scheme. In most cases,

determination of precision and trueness of the measurement procedures is not the primary aim of the PT scheme. To achieve this, further information is often needed and may be obtained from the PT provider.

6.13 Satisfying regulators and accreditation bodies

The successful performance of a laboratory in a PT scheme (or its effective correction of measurement problems after an unsuccessful performance) may provide regulators and accreditation bodies with confidence in the laboratories whose data they endorse or otherwise recognise. The clear benefit for the laboratories is the continuation of their standing as competent organizations.

However, the internal benefits to laboratories, their staff and management, should be of most value if they view PT as a vital tool for ongoing maintenance of confidence and improvement, irrespective of whether the laboratory needs to participate for accreditation purposes.

7 How a PT provider evaluates the laboratory's performance

7.1 Introduction

Results from PT schemes can be in many forms, covering a wide range of data types and underlying statistical distributions. Thus, the purpose of this section is to present the main aspects of the statistical design used by PT providers, so that laboratories can better understand the evaluations performed. This should help the laboratory in the selection of the appropriate PT scheme and in the interpretation of the results. However, given the range of different techniques used it is not possible for this document to address all statistical aspects. It is important that the design used by the PT provider is appropriate for the type and purpose of the PT scheme being organised. Furthermore, the design used by the PT provider should be fully described to the participants. Preferred statistical techniques have been described in ISO 13528 [20], although other valid approaches can be used.

The underlying assumptions of the statistical approach used in PT schemes are mostly based on the normal distribution of data. However, it is common for the set of participant's results, whilst being essentially normally distributed, to show heavy tails and a small proportion of outliers. The original approach used by PT providers (and still used in some PT schemes) was to use statistical tests to identify the presence of outliers in the data set. However, the more common approach now used by PT providers, as recommended in ISO 13528, is to use robust statistics [21, 22]. Robust statistics has the advantage of reducing the contribution of outliers to the calculated statistical parameters such as the mean and standard deviation. There are a number of robust statistical approaches, some of which are described in ISO 13528.

7.2 Basic elements for the evaluation of PT results

7.2.1 General

One of the basic elements in all PT schemes is the evaluation of the performance of each participant. This requires criteria for evaluating reported results. For assessing quantitative results, the PT provider has to establish two values, which are used for the performance evaluation:

1) The assigned value;

2) The SDPA.

In addition, the PT provider would be expected to provide the measurement uncertainty and a statement of the metrological traceability of the assigned value, as stated in ISO/IEC 17043 [5]. The relevance, need and feasibility of the uncertainty evaluation shall be determined by the design of the PT scheme.

Different methods can be used to establish the assigned value and SDPA as described in ISO 13528. There is no strict standardised protocol, which prescribes in detail the statistical design to be used, however this design should be in substantial agreement with the designs described in ISO 13528. The statistical design should be documented by the PT provider, normally either in the scheme protocol or/and in the PT report, and should be taken into consideration by laboratories when selecting an appropriate PT scheme.

NOTE Assessment of qualitative results is considered in 7.2.6

7.2.2 Assigned value

There are, as described in ISO 13528, essentially five methods available to obtain the assigned value, a working estimate of the true value:

1) By formulation;

2) Using a CRM;

3) Results from one laboratory;

4) Consensus value from expert laboratories;

5) Consensus value from participant results.

Further information on these approaches is given in Annex D.

7.2.3 Standard deviation for proficiency assessment (SDPA)

There are, as described in ISO 13528, essentially five approaches to determine the SDPA, i.e. the acceptable range of participant results:

1) By perception of experts;
2) By experience from previous rounds of a proficiency scheme;
3) By use of a general model;
4) Using the repeatability and reproducibility standard deviations from a previous ILC of precision of a measurement method;
5) From data obtained in the same round of a PT scheme.

Further information on these approaches is given in Annex D.

At present a common approach for establishing the assigned value and SDPA is to use the participants' PT results to calculate both values (commonly called "consensus" values). However, it is strongly recommended in the Harmonised Protocol for the Proficiency Testing of Analytical Chemistry Laboratories issued by the IUPAC [23], that scoring methods should be based on fitness for purpose criteria, relevant to the particular circumstances of the determination. Thus, wherever possible, the PT provider should base the SDPA on a fit for purpose value rather than a value that will change from round to round, depending on the spread of the results submitted by the participants. Using a fit for purpose value will facilitate the monitoring of performance scores over successive rounds of the PT scheme.

7.2.4 Performance evaluation

Performance evaluation (or scoring) by the PT provider adds value to the results reported by the participant. The purpose of providing a normalised performance evaluation is to make all PT results comparable, so that the participant can immediately appreciate the significance of the evaluation.

The use of measurement uncertainty in the performance evaluation is increasing as the understanding of this aspect is improving. Two sources of measurement uncertainty must be taken into account:

1) Measurement uncertainty of the assigned value;
2) Measurement uncertainty of the participant's result.

Given the diverse purposes of PT schemes it is not possible to define a single universal evaluation method. Therefore, a number of different methods used for the evaluation of performance are available. The most common are listed below. Other statistical designs, not covered in this document, are given in ISO 13528.

a) "*z* score" is the most commonly used, providing a measure of the deviation of the result from the assigned value. The measurement uncertainties of the assigned value and of the reported result are not taken into account. It is calculated as:

$$z = (x_i - x_{pt}) / \sigma_{pt}$$

where:

x_i is the result reported by participant *i*

x_{pt} is the assigned value

σ_{pt} is the SDPA

b) "*z'* score" is used when there is concern regarding the uncertainty of the assigned value (see 7.2.6), so the standard uncertainty of the assigned value is taken into account:

$$z' = (x_i - x_{pt}) / \sqrt{\sigma_{pt}^2 + u^2(x_{pt})}$$

where:

x_i is the result reported by participant *i*

x_{pt} is the assigned value

σ_{pt} is the SDPA

$u(x_{pt})$ is the standard uncertainty of the assigned value

c) "ζ score" is useful for evaluating a participant's ability to produce results close to the assigned value within their claimed measurement uncertainty. Thus the standard uncertainty of both the assigned value and the participant's result is taken into account:

$$\zeta = (x_i - x_{pt}) / \sqrt{u^2(x_i) + u^2(x_{pt})}$$

where:

x_i is the result reported by participant *i*

x_{pt} is the assigned value

$u(x_i)$ is the standard uncertainty of a result from participant *i*

$u(x_{pt})$ is the standard uncertainty of the assigned value

d) "E_n score" is another scoring system, which takes into account the expanded uncertainty of the assigned value and the participant's result:

$$E_n = (x_i - x_{pt}) / \sqrt{U^2(x_i) + U^2(x_{pt})}$$

where:

x_i is the result reported by participant *i*

x_{pt} is the assigned value

$U(x_i)$ is the expanded uncertainty of reported result from participant *i*

$U(x_{pt})$ is the expanded uncertainty of the assigned value

NOTE ISO 13528 warns that combining expanded uncertainties, as in the E_n score, does not permit a consistent interpretation unless both have the same coverage factor and degrees of freedom.

The following interpretation is commonly used for *z*, *z'* and ζ scores:

i) $|score| \leq 2.0$ the score indicates "satisfactory" performance and generates no signal;

ii) $2.0 < |score| < 3.0$ the score indicates "questionable" performance and generates a warning signal;

iii) $|score| \geq 3.0$ the score indicates "unsatisfactory" performance and generates an action signal.

Some caution should be noted in the interpretation of z' and ζ scores [24]:

- The z' score correctly serves to standardise the deviation from the assigned value, but fails to differentiate between a poor result and a poor assigned value.
- ζ scores increase as either the deviation from the assigned value increases or as the reported uncertainty gets smaller, so a larger ζ score can indicate a large error, an underestimated uncertainty, or both.

The following interpretation is commonly used for E_n scores:

i) $|E_n| \leq 1.0$ the score could indicate "satisfactory" performance and generates no signal if the uncertainties are valid and the deviation (x_i-x_{pt}) is smaller than needed by the participant;

ii) $|E_n| > 1.0$ the score could indicate "unsatisfactory" performance and a need to review the evaluated uncertainty or correct a measurement procedure issue.

The evaluation must be on a consistent basis from round to round of a PT scheme, so that performance scores in successive rounds are comparable. Only in this way can a participant see long-term trends in their performance. It is therefore preferable that the SDPA is based on fitness for purpose criteria (see 7.2.3) rather than based on the spread of participant results.

7.2.5 Alternative performance evaluation approaches

Some PT schemes use a simple difference between assigned value and participant result, often denoted D, as an indication of performance. This can also be expressed as a percentage of the assigned value, D%.

$$D_i = x_i - x_{pt} \qquad D_i\% = 100(x_i - x_{pt})/x_{pt}$$

The difference D or D% is usually compared with a criterion based on fitness for purpose or expected performance. These have the advantage of simplicity for an analyst familiar with the field, but do not have a consistent interpretation for different characteristics.

7.2.6 Effect of the uncertainty of the assigned value

The standard uncertainty of the assigned value depends on a number of factors. These include the method used to derive the assigned value, and when it is derived from measurements made in several laboratories, on the number of laboratories. Methods for calculating the standard uncertainty of the assigned value can be found in ISO 13528.

If the standard uncertainty ($u(x_{pt})$) of the assigned value is too large in comparison with the SDPA, then there is a risk that some laboratories will receive a questionable or unsatisfactory performance because of inaccuracy in the determination of the assigned value, not because of any cause within the laboratories. For this reason, the standard uncertainty of the assigned value is to be established and reported to the laboratories participating in the PT scheme.

If $u(x_{pt}) < 0.3\sigma_{pt}$ [20], then the standard uncertainty of the assigned value is negligible and need not be included in the interpretation of the results of the proficiency test.

If the above criterion is not met, then the PT provider will usually have taken one of the following steps:

a) used a different method for determining the assigned value such that its standard uncertainty meets the above criterion;

b) used the uncertainty of the assigned value in the interpretation of the results of the proficiency test (see above for z' score, ζ score or E_n score);

c) reported separate values and uncertainties for each sub-population (for example, participants using different measurement procedures) if the assigned value was derived from participant results, and the large uncertainty arose from differences between identifiable sub-populations of participants;

d) informed the participants in the proficiency test that the uncertainty of the assigned value is not negligible.

7.2.7 Qualitative and interpretative PT schemes

Qualitative PT schemes (as stated in ISO 13528) and interpretative PT schemes require special consideration for the design, value assignment and performance evaluation (scoring) stages because:

- assigned values are very often based on expert opinion; and
- statistical treatment designed for continuous-valued and count data is not applicable to qualitative data. For example, it is not meaningful to take means and standard deviations of ordinal scale results even when they can be placed in a ranking order.

The following mechanisms for deriving the assigned values in qualitative and interpretative PT schemes are given in ISO 13528:

a) by expert judgement;

b) by use of RMs as proficiency test items;

c) from knowledge of the origin or preparation of the PT item(s);

d) using the mode or median of participant results (the median is not appropriate for nominal values).

Performance evaluation will typically be based on:

- participants judged solely on whether their result coincides exactly with the assigned value for the relevant PT item;
- participants assessed by expert appraisal, which may involve some form of marking system resulting in some type of performance score.

7.2.8 Outliers

An outlier is an observation that is numerically distant from the rest of the data. There is a very low probability that outliers occur by chance. Usually outliers do not belong to the same population of data, i.e. the same distribution. So, they are often indicative of an error in the measurement or in another stage of the analytical process. The PT provider should have either discarded these outliers or used statistics that are robust to outliers. If there is a suspicion that the distribution is non-normal, robust statistical techniques should be used that are robust to asymmetry. A mixture of two or more distributions, which may be two or more distinct sub-populations, e.g. resulting from the use of two different measurement procedures, may sometimes also look as if there were multiple outliers. ISO 13528 gives guidance on the possible consequences.

ISO 13528 specifically addresses the issue of blunder removal. Obvious blunders, such as reporting results in incorrect units or switching results from different PT items, occur in most rounds of PT, and will impair the performance of subsequent statistical methods. It is recommended that PT providers remove obvious blunders from a data set at an early stage in the data analysis, prior to the use of any robust procedures or the application of any tests to identify statistical outliers.

The PT provider should state how it takes outliers into account and how it handles blunders when processing data from PT rounds.

8 Laboratory interpretation of PT results

8.1 Introduction

Taking part in a PT scheme is of limited value unless the laboratory takes advantage of its performance evaluation and the general information given in the PT scheme report.

It is important that the laboratory not only acknowledges the performance evaluation obtained, but evaluates and interprets it, avoiding any misinterpretations or over-interpretations. The evaluation of the performance from the laboratory should be done after each round, and for continuous schemes the performance over time should also be evaluated.

8.2 Performance evaluation by the laboratory

8.2.1 Importance of performance evaluation

The interpretation of the PT performance concerns all management levels of the laboratory, from the operator to the top management. The personnel responsible for the measurement will be familiar with the operation of the PT scheme and should normally proceed with the initial evaluation. If any investigations have to be undertaken, they should be treated within the non-conformity procedure of the laboratory's quality management system. The laboratory management may not always be familiar with PT performance, and it is highly advisable that they gain an appropriate level of understanding of PT schemes.

As the laboratory should be using validated measurement procedures along with IQC's, any poor performance is to be taken seriously as it indicates that there is a problem with the validity of the measurement procedures and/or the IQC procedures.

There are some basic points about the interpretation of PT results, which are worth stating before more detailed consideration of this topic is given. As previously mentioned, PT is not about "passing" or "failing" a measurement; it is about learning from the results. A satisfactory performance in one PT round for a laboratory, where all participants have a satisfactory performance, does not necessarily indicate a high level of competence. In this case it is possible that the SDPA could be too large. Neither, on the other hand, does one unsatisfactory performance in one PT round indicate that the laboratory is not competent; this result needs to be studied and lessons learned from it so that it is not repeated. However, consistent poor performance indicates major problems with the laboratory's measurement procedures and when this occurs the laboratory should give serious consideration as to whether it should continue to offer that particular measurement until the issues are resolved.

8.2.2 Review of results from a single PT round

The results of each PT round are to be evaluated regardless of the performance obtained as a satisfactory result may not necessarily mean a good performance.

All the information available in the PT report should be evaluated, not just the performance score. For example, unsatisfactory performance in the context of a PT round where the majority of participants performed to a satisfactory level should be contrasted with unsatisfactory performance where a significant number of participants had an unsatisfactory performance. Both situations should however be viewed seriously, since both indicate problems regarding the measurement procedures.

As part of the review, the laboratory staff should always check that the results in the PT report are those submitted by the laboratory and, in particular, if the performance scoring system used in the PT scheme is clearly understood and fit for purpose. If necessary, the PT provider should be contacted to avoid any misinterpretation of the performance.

If justified, the laboratory can choose to recalculate its performance score, using a criterion appropriate for their circumstances. For example, a laboratory's customers might not require the level of performance implied by the scheme's usual SDPA, or the scheme SDPA may have been based on the standard deviation of an unusually similar set of participant results (see 8.2.3).

If after a thorough investigation, the laboratory concludes that the result is indeed unsatisfactory, then corrective actions should be initiated (see section 8.3).

The laboratory's results from its participation in PT can also be used to check the validity of the laboratory's measurement uncertainty (see section 6 and 7.2.4 c)).

8.2.3 Evaluation of the fitness for purpose of *z* score calculations

For performance evaluation, the z score (and the related z' and ζ scores) are most commonly used. The z score is calculated by dividing the difference between the the laboratory's test result and the assigned value by the SDPA. The assessment of whether a laboratory's result is considered satisfactory, questionable or unsatisfactory therefore strongly depends on the value of the SDPA used in the denominator of the z score equation. The bigger the SDPA the smaller the z score and, therefore, the better the apparent performance of the laboratory in the PT round. The question the laboratory has to answer is whether the SDPA has been chosen on a sound basis and whether it is fit for the laboratory's purpose. Issues to consider include:

a) Is the SDPA derived on a purely statistical basis?
   In this case the estimate of the SDPA becomes unreliable with a small number of participants, a wide spread of results, or both. A small number of participants may therefore require the SDPA to be set using an alternative approach (see section 7.2.3 and Annex D). The same applies if the SDPA is derived from statistical data from historic rounds of the PT scheme, if these PT rounds have a small number of participants, a wide spread of results, or both.

b) Is the SDPA derived from or based on legislative or normative documents?
   Legislation or standards often contain information on the repeatability and/or reproducibility of measurement procedures. This information can be used to judge if the SDPA is fit for purpose.

c) Is the SDPA consistent with the laboratory's own QA data?
   Estimates of repeatability and reproducibility derived from validation or verification studies, results from the analysis of QC materials and information on the SDPA used by other PT providers can give an indication of whether the SDPA is realistic.

d) Is the SDPA realistic compared to market requirements?
   The requirements from the customer for repeatability and measurement uncertainty regarding the results or for conditions for analyses can be used as indicators as well as requirements for the accuracy for measurement procedures derived from fail and pass decisions for products.

If the laboratory comes to the conclusion that the SDPA is too small or big for its purpose, it should consider choosing an alternative value and recalculating its performance score. If the laboratory uses a smaller SDPA, this needs to be used with caution in that this would only be possible if the level of homogeneity and stability of the PT items were still appropriate [20]. In most cases this will require a discussion with the PT provider since the information required to be able to check this will not be available in the PT report. If the level of homogeneity and/or stability is not sufficient, or cannot be checked, the PT is not fit for purpose and the laboratory should choose another PT which fits with their requirements.

8.2.4 Monitoring PT performance over time

In addition to the careful evaluation of results from individual PT rounds, the performance over time should be monitored, in order to identify potential problems related to imprecision, systematic error or human error. There are a variety of ways that an individual laboratory can monitor its performance over time, such as using graphical approaches or by the calculation of long-term bias components [25, 26].

A plot of performance scores from PT round to round in order to monitor laboratory performance is very useful. This is often given by the PT provider in the PT report, or can be plotted by the participant. This approach enables unusual or unexpected results to be highlighted, as well as assisting in the identification of trends. A laboratory's IQC procedures would normally be expected to identify trends associated with, for example, improper instrumental calibration or maintenance, or mishandling/inappropriate use of reagents. Monitoring PT performance over time complements routine IQC procedures. Examples of typical plots can be found in ISO 13528.

To be able to determine whether a laboratory's performance is improving or deteriorating over time, the data from subsequent PT rounds has to be comparable. However, if the SDPA is obtained from the data set for each PT round, the same measurement from two different PT rounds may have a different SDPA and so lead to the performance scores being calculated differently. If the SDPA used in consecutive PT rounds differs substantially, the participants could calculate their own z scores (or other performance scores) using a suitable SDPA. A SDPA obtained from literature (e.g. from a standard measurement procedure published by a national or international standardisation body such as ISO or DIN) can be used. If such values are not available, the laboratory may set its own criteria based on the purpose of participating in the PT scheme or the importance of the measurement. The laboratory can choose any suitable value as long as it can justify its choice. Note that the selected SDPA does not have to be constant, i.e. it may be concentration dependent. Some caution needs to be given when selecting a different SDPA (see 8.2.3) from that used by the PT provider. If a laboratory decides to recalculate their performance scores, they should justify and document their choice.

The use of combined performance scores (for example averaged or summed on the same or different PT items) should be used only with caution [20]. Where provided by a PT provider the limitations of such an approach should have been made clear in the PT report.

8.3 Investigation of unsatisfactory or questionable PT results

8.3.1 Need for an investigation

All laboratories will occasionally have unsatisfactory or questionable PT results. When this occurs, the laboratory should clearly identify and document them.

The depth of the investigation that has to be undertaken will depend upon a number of factors. These include, the criticality of the measurement procedure, the frequency of unsatisfactory results and evidence of a bias. In every case, the laboratory should document the evaluation of the results, even if it decides not to take any specific action.

As a basic principle, every unsatisfactory performance score should be investigated and the investigation documented as this clearly denotes a problem. The laboratory should have a policy on when to investigate:

a) Questionable performance scores for the same measurement;

b) Consecutive performance scores, for the same measurement, which have the same bias sign against the assigned value.

However, it is important to note that it is up to the laboratory to specify its own criteria [2, 3] for launching an investigation, taking into consideration the frequency of participation, the fitness of purpose of the PT scheme, the criticality of the measurement, etc. The key issue is that unsatisfactory performance needs to be investigated and trends should be examined.

8.3.2 Root cause investigation

When a full investigation is deemed necessary, a stepwise approach is preferred, in order to maximize the chances of determining the root cause of the problem. A diagram supporting this approach is given in Annex B.

An adequate stepwise investigation procedure should consist of the following steps and involve the personnel that performed the measurement:

a) analyse the problem based on the raw data, the overall performance of the PT round, the results from successive PT rounds and IQC data;

b) make a plan for corrective action(s);

c) execute and record the corrective action(s);

d) check whether the corrective action(s) was effective.

8.3.3 Causes of poor performance

8.3.3.1 Typical causes

The reasons for obtaining a poor performance are unfortunately numerous, potentially resulting in a time consuming and complex investigation. However, as the investigations should result in an improvement of the laboratory's performance, it is worthwhile to put in the necessary effort. In order to facilitate the investigations, it is useful to have in mind the main causes for poor performance so that the investigations can be better focused. Typical causes of poor performance include [27]:

a) sample preparation (e.g. weighing, drying, extraction, digestion, clean-up, dilution, etc.);

b) measurement procedures;

c) human error (e.g. inappropriate training, transcription errors);

d) calibration;

e) selection of measurement procedure;

f) calculation error;

g) reporting problem (e.g. format, unit, interpretation);

h) problem arising from the PT item;

i) sample transport and storage;

j) primary sampling;

k) sample tracking (e.g. labelling, chain of custody);

l) problem arising within the PT provider.

In order to identify the root cause of poor performance, it is important to focus on the potential causes, which can be grouped as:

i) clerical error;

ii) technical problem (e.g. measurement procedure, equipment, training, IQCs);

iii) problem related to the PT scheme (e.g. inadequate PT scheme, inappropriate evaluation).

It is possible that after a thorough investigation, the origin of the poor performance is not identified. If it is a repeated poor performance, then the laboratory procedures (technical or management) should be questioned.

8.3.3.2 Clerical error

Although clerical errors are not directly linked to the laboratory's technical competence, they can underline that the laboratory may have a potential problem when reporting results to the customers.

Clerical errors can include the following:

a) transcription errors;

b) mislabelling;

c) decimal error;

d) results reported in the wrong units.

Identifying if a clerical error has been made is an important first step of an investigation. If clerical errors are a regular cause of unsatisfactory results, then the investigation should be focused on the quality aspects of the management system.

8.3.3.3 Technical problem

Due to the complexity of laboratory activity, problems can occur at every level of the laboratory procedures and each of the following elements should be reviewed during the investigation:

a) storage/pre-treatment of the PT item;

b) measurement procedure/IQC data;

c) equipment/reagents/calibration;

d) environmental conditions;

e) data processing.

If the investigation of the laboratory procedures does not enable the laboratory to identify the root cause, it may be necessary to review the validation of the measurement procedure.

8.3.3.4 Problem related to the PT scheme

Poor performance could also be due to the fact that the selected PT scheme was inappropriate or that a problem occurred with the PT items. The following points should be investigated:

a) matrix difference between PT item and routine samples;

b) potential PT item deterioration;

c) concentration levels outside the scope of application of the measurement procedure;

d) lack of stability or homogeneity of the PT items;

e) inappropriate instructions to participants;

f) PT item storage problems;

g) inappropriate peer group;

h) inappropriate assigned value;

i) inappropriate SDPA;

j) incorrect data entry by the PT provider.

The laboratory is encouraged to discuss their findings with the PT provider or they may wish to evaluate if the PT scheme selected is appropriate.

Annex A: Selecting the most relevant PT scheme

PT Item

- What is the matrix?
- Is the PT item real or simulated?
- Are all the characteristics routinely tested available?
- Are the chracteristic values (e.g. concentrations) appropriate?
- Are standard reporting units used?

Participants

- Is the participant base national or international?
- Is the number of participants or the size of the peer group appropriate?
- What measurement procedures are being used by participants?
- What type of laboratories are participating?

PT item distribution

- Are the distribution dates available and appropriate?
- Does the frequency of distributions meet the needs of the laboratory?
- Does the PT provider allow flexible participation?

Results

- Are the deadlines for submitting results available and appropriate?
- How does the PT provider require the results to be reported?
- Can participants report results obtained using their choice of measurement procedure?
- Can measurement uncertainties be reported and will they be included in the performance evaluation?
- Is the statistical approach used available and appropriate?

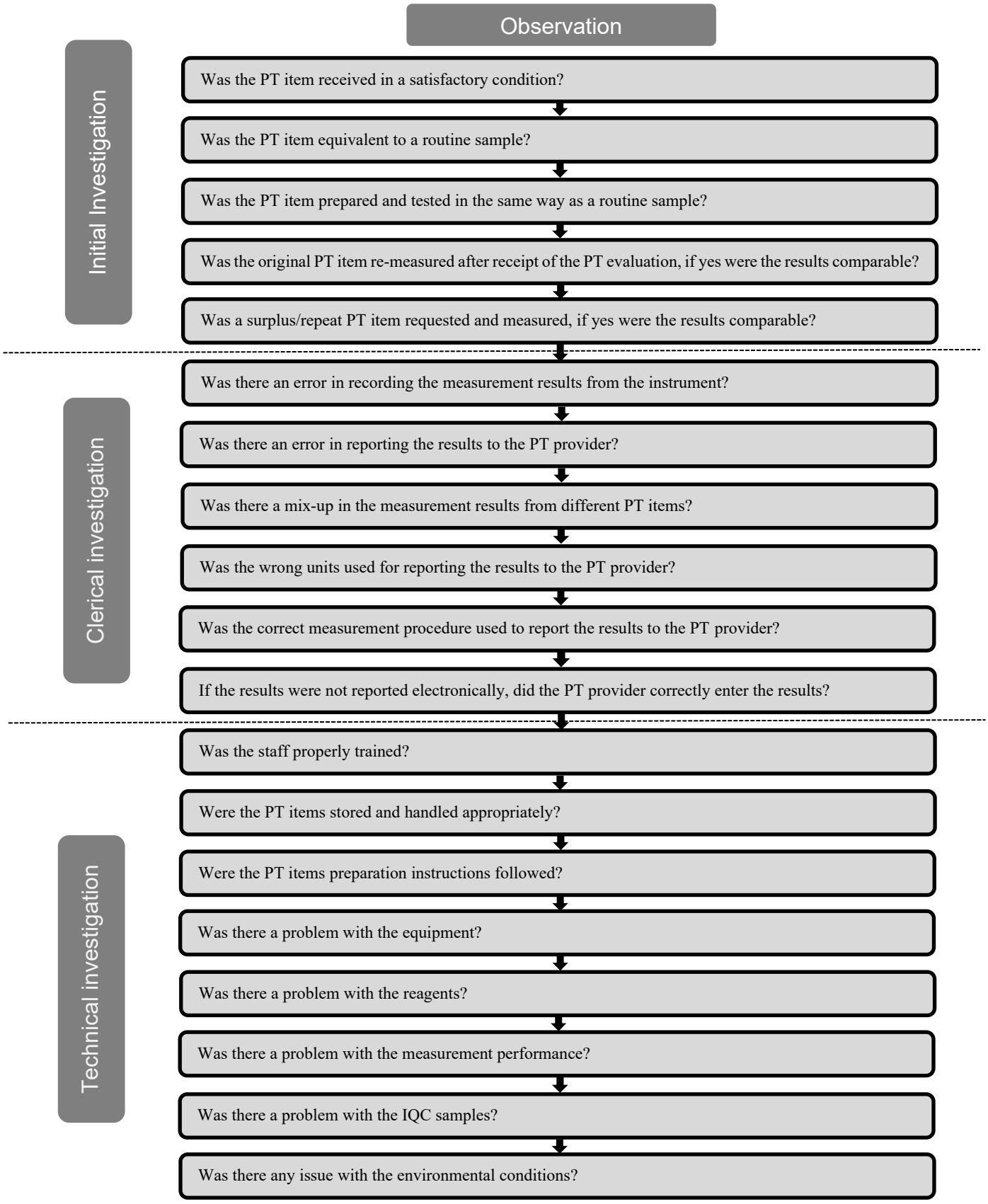
PT reports

- How quickly after the result deadline are PT reports provided?
- What information is provided in the PT reports?
- Are the evaluation criteria (e.g. SDPA) fit for the laboratory's purpose?
- Is the PT report available on paper, electronically or online?
- Does the report include easily interpretable graphical summaries?
- Is the language is used in the PT reports understood by the relevany staff?

PT providers

- What is the scope of PT schemes offered?
- Is appropriate feedback and assistance provided?
- Are "surplus/repeat PT items" provided to laboratories for carrying out investigations of poor performance and evaluation of the effectiveness of corrective actions?
- Do they comply with the requirements of ISO/IEC 17043?
- Are they accredited to ISO/IEC 17043 by an accreditation body?

Annex B: Investigating unsatisfactory or questionable PT results



Annex C: Interpretation of PT data by end-users

C.1 Introduction

Laboratories will need to demonstrate their competence to interested parties such as accreditation bodies, regulatory bodies and customers. PT results, as well as the other QC activities are some of the means to demonstrate competence. As PT is usually a third-party evaluation, the interested parties are increasingly recommending or requiring participation of laboratories in PT schemes in order to have an independent evaluation of the performance of the laboratory. Participation in PT is a mechanism for external QC of results and for comparison with other laboratories as required in the standards ISO/IEC 17025 [2] and ISO 15189 [3].

It is the responsibility of the laboratory to ensure that when providing its PT results to interested parties, they also provide all the appropriate additional information (e.g. recalculated performance score, investigations).

C.2 Accreditation bodies

Accreditation bodies, and technical assessors employed by accreditation bodies, generally have a good understanding of the role of PT, and are skilled in the interpretation of PT scheme results obtained by laboratories that are either accredited or seeking accreditation. In general, the technical assessors are familiar with PT schemes in which the laboratory participates. PT scheme protocols and other documentation will be studied and, if necessary, the PT provider contacted, by the accreditation body, to discuss or clarify any outstanding issues. The level of performance in a PT scheme for any laboratory will be determined against the criteria established by the PT provider. In some cases, what constitutes unsatisfactory performance within a PT scheme may still be acceptable or fit for purpose within the scope of the laboratory's accreditation and vice-versa.

C.3 Regulatory bodies

Regulatory bodies have the need to ensure that measurements made in laboratories that are covered by regulations or directives are of satisfactory quality. Therefore, regulatory bodies may use PT scheme performance as one of the ways of assessing quality in addition to other approaches including having referee analyses undertaken or submitting check samples for analysis.

Where a regulatory body has been involved in the development of a PT scheme, it will incorporate features that are of direct relevance to that body, and will be readily understood. For those situations where the regulatory body is using an independent PT scheme for their own purposes, it is recommended that they discuss fully the scope and operational parameters of the PT scheme with the PT provider. This will enable them to put results obtained by any laboratory of interest into context. The statistical processes used by the PT provider for the calculation of laboratory performance needs to be understood, in order that a laboratory's performance may be judged in relation to any tolerances allowed in regulations. Advice may be required from the PT provider in such situations in order that PT scheme performance data is not misinterpreted.

C.4 Customers of participant laboratories

The customer of a laboratory participating in a PT scheme can use the performance in the PT scheme as one tool with which to monitor the quality of that laboratory. The customer needs to have a good understanding of how the PT scheme operates and how the PT provider calculates performance within the PT scheme. Although some systems for determining performance in a PT scheme are widespread, such as the use of the z score, there are many different systems in use. In addition, customers should be aware that the way in which z scores and other performance indicators are calculated can vary between PT schemes.

Customers are increasingly including PT scheme performance criteria in tender documents, and are using information about PT scheme performance supplied by potential contractors to assist in the decision as to which laboratory is awarded the contract. When using PT scheme performance as a criterion in a tender, customers should ensure that, where they are setting a "performance standard", it is realistic and achievable. For example, asking laboratories to achieve satisfactory results for all characteristics in all PT rounds of a PT scheme is unrealistic. PT providers normally provide appropriate information on the overall performance of the PT scheme in the PT report, so that a good

benchmark may be set. Customers should also take care to ensure that the matrices/characteristics in which they have an interest are clearly stated, as the PT scheme may have a broader scope, and performance of laboratories for matrices/characteristics not of direct interest may be irrelevant.

Customers must place any data relating to PT scheme performance from a contract laboratory into the proper context; laboratories could present data to a customer in a way that paints an unrealistically positive picture.

Customers are recommended to carry out the following, as appropriate, in order to gain an accurate picture of the laboratory's true performance:

a) obtain information on the scope and operation of the PT scheme (e.g. PT scheme protocol) from the laboratory or the PT provider;

b) look at laboratory performance over time, since one PT round in a PT scheme only gives a brief snapshot of the laboratory's performance;

c) review the overall performance of all participants in order to judge how the laboratory is performing;

d) ask for copies of PT scheme reports (where confidentiality is not an issue) to confirm any data summarising PT scheme performance. The PT provider may provide this data, although the agreement of the participant will generally be required.

One unsatisfactory result in any PT round of a PT scheme does not make a laboratory poor, neither does the achievement of 100 % satisfactory results in any PT round make a laboratory necessarily good.

The way in which a laboratory responds to an unsatisfactory result will usually give more information about that laboratory than the occurrence of the unsatisfactory result.

Annex D: Statistical aspects for PT

D.1. Main parameters

One of the basic elements in all PT is the evaluation of the performance of each participant. In order to do so, the PT provider has to establish two values, which are used for the performance evaluation:

1) The assigned value;
2) The SDPA.

As mentioned in section 7.2, the PT provider has to also evaluate the measurement uncertainty of the assigned value.

D.2 Assigned value and standard uncertainty of the assigned value

There are, as described in ISO 13528 [20], essentially five methods available to obtain the assigned value and its associated standard uncertainty. The choice of method is the responsibility of the PT provider:

1) Formulation: the mixing of materials with different known concentration levels in specified proportions, or the addition of a known amount or concentration of an analyte to a base material containing none. This method is satisfactory in many cases, especially when it is the total amount of the analyte rather than a concentration that is subject to measurement but, of course, it may not simulate the difficulty of normal sample preparation procedures (which may include steps such as extraction and speciation) where recovery problems may well arise.

   The standard uncertainty is evaluated by combining the uncertainties associated with the preparation of the PT item using an appropriate model.

2) CRM: when the PT item is a CRM, its certified value is used as the assigned value. This has the advantage of providing a traceable assigned value, but it is an expensive approach and appropriate CRMs are often not available. Furthermore, CRMs are often highly processed to ensure long-term stability, which may compromise the commutability [28] of the PT items.

   The standard uncertainty is derived from the information on uncertainty provided on the certificate for the CRM.

3) Results from one laboratory: a selection of the prepared PT items is measured, by a chosen laboratory, either using a primary method or alongside a CRM. The assigned value is derived directly from the primary method used or from a calibration against the certified value of the CRM. This will provide a traceable assigned value via the primary method or to the CRM used, but it relies on the results from a single laboratory and appropriate primary methods or CRMs may not be available. The reference method, or the CRM used as a reference, should be commutable for all the measurement methods used by participants [28].

   The standard uncertainty is derived from the measurement results of the chosen laboratory, and the uncertainties of the certified values of the CRM.

4) Consensus value from expert laboratories: the determination of a consensus value obtained from the outcome of a group of expert laboratories being proficient in the measurement procedures applied. However, it is often hard or even impossible to find a group of expert laboratories whose expertise is beyond doubt and accepted by all participants of the PT. This is even more true for large, international PT schemes with participants from many countries. For a number of measurands (for example, extractable lead in soil), the true value of the parameter being measured is, in principle, defined by the measurement procedure used. In such cases, the definition ‘operationally defined measurands’ is often used and the measurement procedures (methods) are referred to as ‘empirical methods’. In these cases, the expert laboratories should all use the same measurement procedure and should follow it in every detail. There may be an unknown bias in the results of the group of expert laboratories. The expert laboratories and the measurement procedures applied should be declared before the PT scheme is set up.

   Where the expert laboratories report uncertainties with the results, the estimation of a value by consensus of results is a complex problem and a wide variety of approaches has been suggested, including, for example, weighted averages, un-weighted averages, procedures that make allowance for over dispersion and procedures that allow

for possible outlying or erroneous results and evaluated uncertainties. The PT provider should accordingly establish a procedure for estimation that:

a) includes checks for validity of reported uncertainties, for example by checking whether reported uncertainties account fully for the observed dispersion of results;

b) uses a weighting procedure appropriate for the scale and reliability of the reported uncertainties, which may include equal weighting if the reported uncertainties are either similar or of poor or unknown reliability;

c) allows for the possibility that reported uncertainties might not account fully for the observed dispersion ('over dispersion'), for example by including an additional term to allow for over dispersion;

d) allows for the possibility of unexpected outlying values for the reported result or the uncertainty;

e) has a sound theoretical basis;

f) has been verified, for example on test data or in simulations, to show that it is sufficient for the purposes of the PT scheme.

5) Consensus value from participant results: the use of a consensus value, produced in each round of the PT scheme, and based on the results obtained by the participants. The consensus value is usually estimated using robust statistical techniques. The consensus approach is clearly the most straightforward and in some cases, for example, when using natural matrix PT items, may be the only way to establish an estimate of the true value.

   A common estimate of the uncertainty for a consensus assigned value obtained by a robust statistical procedure is:

$$u(x_{pt}) = 1.253 \times \left(\frac{s^*}{\sqrt{p}}\right)$$

   where:

   s^* is the robust estimate of the participant standard deviation

   p is the number of participants

   The factor of 1.25 is based on the variance of the median for normally distributed data.

   Other valid approaches are described in, for example, ISO 13528.

   The limitations of this approach are that:

   a) there may be no real consensus amongst the participants;

   b) the consensus may be biased by the general use of faulty measurement procedures and this bias will not be reflected in the standard uncertainty of the assigned value calculated as described above.

D.3 SDPA

There are, as described in ISO 13528, essentially five approaches to determine the SDPA, i.e. the acceptable range of participant results:

1) By perception of experts: the SDPA may be set at a value that corresponds to the level of performance that a regulatory authority, accreditation body, or the technical experts of the PT provider believe is reasonable for participants.

2) By experience from previous rounds of a PT scheme: the SDPA may be set at a value that is based on the experience of previous rounds of the PT scheme. This corresponds to the level of performance that the PT provider would wish laboratories to be able to achieve.

3) By use of a general model: the value of the SDPA may be derived from a general model for the reproducibility of the measurement procedure, such as a concentration dependent model. This method has the advantage of objectivity and consistency across characteristics being measured, as well as being empirically based.

A disadvantage of this approach is that the true reproducibility of a particular measurement procedure may differ substantially from the value given by the model as the use of a general model implies that the reproducibility depends only on the concentration level of the analyte, and not on the analyte, the measurement procedure, or the sample size.

4) Using the repeatability and reproducibility standard deviations obtained from a previous ILC carried out to assess the performance characteristics of a measurement procedure: when the measurement procedure to be used in the PT scheme is standardized, and information on the repeatability and reproducibility of the measurement procedure is available, the SDPA may be calculated using this information.

5) From data obtained in the same round of a PT scheme: with this approach, the SDPA used in a round of a PT scheme is derived from the results reported by the participants in the same PT round. It shall be the robust standard deviation of the results reported by all the participants.

   A disadvantage of this approach is that the value may vary substantially from PT round to round, making it difficult for a laboratory to use its *z* score to look for trends that persist over several PT rounds.

Bibliography

For update of current most important references please refer to the Eurachem Reading List placed under Publications at the Eurachem website, www.eurachem.org.

1. EA-4/21 INF:2018, Guidelines for the assessment of the appropriateness of small interlaboratory comparisons within the process of laboratory accreditation, available from www.european-accreditation.org
2. ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories, ISO, Geneva
3. ISO 15189:2012, Medical laboratories – Requirements for quality and competence, ISO, Geneva
4. Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, CIPM, October 1999, available from www.bipm.org/en/cipm-mra
5. ISO/IEC 17043:2010, Conformity assessment – General requirements for proficiency testing, ISO, Geneva
6. JCGM 200:2012, International vocabulary of metrology — Basic and general concepts and associated terms (VIM, 3rd edition), available from www.bipm.org/en/publications
7. EA-4/18 G:2021, Guidance on the level and frequency of proficiency testing participation, available from www.european-accreditation.org
8. Eurachem - Information leaflet on pre- and post-analytical proficiency testing, First English edition, 2009-05-14, available from www.eurachem.org
9. EPTIS PT scheme database, www.eptis.org
10. IFCC laboratory medicine PT database, available from www.ifcc.org/ifcc-scientific-division/sd-committees/cmd/externalqualityassessment-proficiencytestinginmoleculardiagnostics
11. Sandberg, S. et al., Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem Lab Med 2015:53(6):833-835
12. ILAC Brochure: 2019, Benefits for laboratories participating in proficiency testing programs, available from www.ilac.org/publications-and-resources/ilac-promotional-brochures
13. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU
14. Eurolab Technical Report 1/2007 – Measurement uncertainty revisited, Alternative approaches to uncertainty evaluation, March 2007, available from www.eurolab.org/pubs-techreports
15. ISO 21748:2017, Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation, ISO, Geneva
16. Ellison, S.L.R and Williams, A. (eds), Eurachem/CITAC Guide: Quantifying uncertainty in Analytical Measurement, 3rd Edition, 2012, available from www.eurachem.org
17. NORDTEST Technical Report 537: Handbook for calculation of measurement uncertainty in environmental laboratories, NORDTEST 2012, available from www.nordtest.info/index.php/technical-reports/category/environment
18. Ramsey, M. H., Ellison, S. L. R. and Rostron, P. (eds.) Eurachem/EUROLAB/ CITAC/Nordtest/AMC Guide: Measurement uncertainty arising from sampling: a guide to methods and approaches. Second Edition, Eurachem (2019). ISBN (978-0-948926-35-8), available from www.eurachem.org
19. Eurachem - Information leaflet on use of surplus proficiency test items, First English edition, November 2019, available from www.eurachem.org
20. ISO 13528:2015, Statistical methods for use in proficiency testing by interlaboratory comparison, ISO, Geneva

21. Analytical Methods Committee – Robust Statistics Part I & II. Analyst 1989:114:1693-1702

22. Thompson, M. and Ellison, S.L.R., Fitness for purpose – the integrating theme of the revised Harmonised Protocol for Proficiency Testing in Analytical Chemistry Laboratories. Accred. Qual. Assur. 2006:11:373-378

23. Thompson, M., Ellison, S.L.R. and Wood, R., The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure Appl. Chem. 2006:78:1:145-196

24. AMC technical brief No. 74: Z-Scores and other scores in chemical proficiency testing – their meanings, and some common misconceptions, Anal. Methods 2016:8:5553

25. Meijer, P., DE Maat, M.P.M, Kluft, C., Haverkate, F., and van Houwelingen, H.C., Long-Term Analytical Performance of Hemostasis Field Methods as Assessed by Evaluation of the Results of an External Quality Assessment Program for Antithrombin, Clin. Chem. 2002: 48:7:1011-1015

26. Meijer, P., Haverkate, F., Kluft, C., Performance goals for the laboratory testing of antithrombin, protein C and protein *S*. Thromb. Haemost. 2006:96:5:584-589

27. Ellison, S.L.R. and Hardcastle, W.A., Causes of error in analytical chemistry: results of a web-based survey of proficiency testing participants, Accred. Qual Assur. 2012: 17, 453–464, available from https://doi.org/10.1007/s00769-012-0894-2

28. ISO Guide 35:2017, Reference materials – Guidance for characterization and assessment of homogeneity and stability, ISO, Geneva