Eurachem

A focus for analytical chemistry in Europe

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

Second Edition 2011

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

Second Edition 2011

Editors

Ian Mann (SAS, Switzerland)
Brian Brookman (LGC Standards, UK)

Composition of EEE-PT Working Group

Brian Brookman (Chair), LGC Standards, UK
Natalie Stephenson (Secretary), LGC Standards, UK
Frank Baumeister, AQS Baden-Württemberg, Germany
Greg Ewbank, UKAS, UK
Holger Frenz, Institut für Eignungsprüfung (IfEP), Germany*
Maria Grazia del Monte, ACCREDIA, Italy
Magnus Holmgren, SP Technical Research Institute, Sweden
Eva Klokocnikova, CAI, Czech Republic
Christian Lehmann, DAkkS, Germany
Ulrich Leist, DRRR, Germany
Mirja Leivuori, SYKE, Finland
Minna Loikkanen, Labquality, Finland
Ian Mann, SAS, Switzerland*
Piet Meijer, ECAT Foundation, The Netherlands*
Marina Patriarca, Istituto Superiore di Sanità (ISS), Italy
Poul Erik Poulsen, DANAK, Denmark
Piotr Robouch, IRMM, European Commission*
Pedro Rosario Ruiz, RPS-QUALITAS, Spain*
Annette Thomas, WEQAS, UK
Johannes van de Kreeke, BAM,Germany

*Members of the drafting group led by Ian Mann

Acknowledgements

The revision of this document was undertaken 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.

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

English edition

2nd edition 2011

Enquiries regarding the translation, and production and distribution of new editions of this document should be directed to the EURACHEM Secretariat

CONTENTS

1 INTRODUCTION 1

2 SCOPE 2

3 DEFINITIONS 2

4 INTRODUCTION TO PROFICIENCY TESTING 4

4.1 Role of PT within the management system 4
4.2 Types of PT schemes 5

5 SELECTION OF PT SCHEMES 6

5.1 Strategy of PT participation 7
5.2 Availability of PT schemes 8
5.3 How to decide if the selected PT scheme is fit for purpose? 8

6 USE OF PT BY LABORATORIES 9

7 HOW A PT PROVIDER EVALUATES THE LABORATORY'S PERFORMANCE 12

7.1 Introduction 12
7.2 Basic elements for the evaluation of PT results 12
7.2.1 Assigned value 13
7.2.2 Standard deviation for proficiency assessment 13
7.2.3 Performance evaluation 14
7.2.4 Effect of the uncertainty of the assigned value 15
7.2.5 Qualitative PT schemes 15
7.2.6 Outliers 16

8 LABORATORY INTERPRETATION OF PT RESULTS 16

8.1 Performance evaluation by the laboratory 16
8.1.1 Review of single PT round results 17
8.1.2 Monitoring PT performance over time 17
8.2 Investigation of unsatisfactory or questionable PT results 18
8.2.1 Need for an investigation 18
8.2.2 Root cause investigation 18
8.2.3 Causes for poor performance 19

Appendix A: Selection form for a relevant PT scheme 21

Appendix B: Form for documenting PT investigations 22

Appendix C: Interpretation of PT data by end-users 24

Appendix D: Statistical aspects for PT 26

Appendix E: Performance evaluation for PT 29

Appendix F: Examples of different performance evaluation approaches 32

Appendix G: Example of long-term performance evaluation 38

Appendix H: Example of the use of measurement uncertainty 41

Bibliography 45

1 INTRODUCTION

A regular independent assessment of the technical performance of a laboratory is necessary to assure the validity of measurements or tests (the term “measurement” is used in this document and covers both measurement and tests), 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. 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 results with those of other laboratories. This is achieved by the PT scheme provider distributing homogeneous and stable PT items to participants for analysis and reporting of the 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 accuracy [1] of its measurements. In addition, participation in an appropriate PT scheme is required for laboratories seeking recognition of their competence through accreditation against the standard 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 the laboratory results. It is important for PT scheme providers to bear in mind the needs of these organisations 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 Appendix 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 scheme(s) they should participate in. It is important that this type of information is widely available in order to be able to select the most appropriate scheme.

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

There are a number of key principles, covered in this document, which help to ensure the appropriateness of participation in PT schemes that 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 in terms of measurement samples, analytes and levels; any differences should be noted and accounted for;

b) laboratories should treat PT items as routine samples, i.e. not give them special treatment;

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

d) 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 and protocols must provide clear information in order for all parties to understand how the scheme operates;

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

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

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 focuses mainly on quantitative PT schemes, as it is the type of PT scheme that is most used by the laboratories. Nevertheless, there is also some guidance on the qualitative and interpretive schemes. Many of the general principles are however applicable for all types of PT schemes.

Although this document is primarily aimed at testing laboratories, some of the principles mentioned in this document may apply to calibration laboratories. The information can also apply to other participants in PT schemes such as individuals, organisations or inspection bodies.
This document does not address those ILCs that are aimed at the evaluation of performance characteristics of a method or the assignment of values to reference materials as well as the “key comparisons” that are aimed for the National Metrology Institutes.

The information can also be very useful for other parties such as accreditation bodies, regulatory authorities or customers of the laboratory.

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] [4]

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 (see ISO/IEC 17043) for their proficiency testing schemes and/or for their broader programmes.

[ISO/IEC 17043, definition 3.7] [4]

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] [4]

3.4: 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] [4]

3.5: assigned value
value attributed to a particular property of a proficiency test item

[ISO/IEC 17043, definition 3.1] [4]

3.6: standard deviation for proficiency assessment
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] [4]

3.7: 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 and includes 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.

[ISO/IEC Guide 99:2007 [5] /JCGM 200:2008, definition 2.1] [5]

3.8: 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.

[ISO/IEC Guide 99:2007 [5] / [JCGM 200:2008, definition 2.26] [5]

3.9: measurement technique
the process of testing/calibrating/identifying the property, including any pre-treatment required to present the sample, as received by the laboratory, to the measuring device (e.g. ICP-MS, Rockwell Hardness, PCR, Microscopy, Force Measurement)

[EA-4/18] [6]

3.10: property
the quantity being measured (e.g. Arsenic, Creatinine, Length, Hardness, Force)

[EA-4/18] [6]

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

3.11: product
the item that the measurement technique is being applied to (e.g. Soil, Vegetables, Serum, Polystyrene, Concrete)

[EA-4/18] [6]

3.12: level of participation
the number of sub-disciplines that an organisation identifies within its scope, and therefore the number of specific proficiency tests that should be considered for participation

[EA-4/18] [6]

3.13: frequency of participation
this is how often a laboratory determines that it needs to participate in PT for a given sub-discipline, this may vary from sub-discipline to sub-discipline within a laboratory and between laboratories with the same sub-disciplines

[EA-4/18] [6]

3.14: Sub-discipline (area of technical competence)
an area of technical competence defined by a minimum of one Measurement Technique, Property and Product, which are related (e.g. Determination of Arsenic in soil by ICP-MS)

[EA-4/18] [6]

4 INTRODUCTION TO PROFICIENCY TESTING

4.1 Role of PT within the management system

In order to monitor the reliability of its measurements, it is important for the laboratory to implement quality control measures. For laboratories that are accredited, or seeking accreditation, these measures are an important aspect of the requirements. PT is one of these measures, as well as for example, the use of (certified) reference materials.

PT plays a highly valuable role as it provides an objective evidence 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 does not only give information on the performance of the analytical system, but also on other aspects of the management system such as reception/treatment of the sample, treatment of the data, result reporting etc. It is most important that the laboratory sets up a relevant strategy for participation in PT schemes (see also 5.1).

PT provides an opportunity to also undertake comparisons of the participant's data compared to assigned values (or other performance criteria) or to the performance of their peers. The results from such participation will provide participants with either a confirmation that their performance is satisfactory or an alert that investigation of potential problems is required.

Human and financial resources needed for PT participation, can be important, but should not be limited to such a degree that there are risks that a participant's data may have errors, biases or significant differences compared to their peers that remain undetected.

It is important to underline that the aim of PT participation is not just about performing well or badly, 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 categories:

1. a) qualitative: the results of qualitative tests are descriptive and reported on a nominal or ordinal scale;

   NOTE 1 A nominal scale is a numerical system used for classifying and categorizing data, e.g. gender (Male/female). A nominal scale is also sometimes called a categorical scale

   NOTE 2 An ordinal scale is a measurement scale that assigns values to objects based on their ranking with respect to one another, e.g. +, ++, +++.

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

   NOTE 3 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 thing being measured, e.g. Fahrenheit and Celsius temperature scales.

   NOTE 4 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 thing being measured, e.g. a "4" on such a scale implies twice as much of the thing being measured as a "2".

   c) interpretive: no measurement is involved. The PT item is a measurement result, a set of data or other set of information concerning an interpretative feature of the participant's competence;

2. a) single: PT items are provided on a single occasion;

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

3. a) sequential: 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;

   b) 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. a) pre-measurement [7]: in this type of PT scheme, the “PT item” can be an item (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);

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

   c) post-measurement: 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.

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 of the types of PT schemes mentioned above could be organised as a blind PT.

5 SELECTION OF PT SCHEMES

The selection of a PT scheme is critical to ensure that the participant obtains the most benefit from participating; therefore the selection process of an appropriate PT is important. It is therefore essential that the participant evaluates the competence of the PT provider. A PT provider that operates according to ISO/IEC 17043 [4] can be considered as competent.

Participating in a PT scheme provides a laboratory with an objective means of assessing and demonstrating the reliability of the results it produces. It thus supplements a laboratory’s own internal quality control 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.

In selecting the appropriate PT scheme, within an area of technical competence, a laboratory should answer the following questions:

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, i.e. does the PT provider operate according to ISO/IEC 17043 [4]?

5) Is the selected PT scheme independent of any manufacturing or marketing interests in equipment, reagents or calibrators in its field of operation?

5.1 Strategy of PT participation

Before selecting a PT scheme, laboratories should evaluate the level and frequency of its 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.

In cases where PTs are required as mandatory by the regulatory authorities, the laboratory has no choice to select the most appropriate scheme and frequency. But it can still consider them in its PT participation strategy to cover some of the fields where PT participation was decided to be useful.

The areas of technical competence (or sub-discipline) of a laboratory can be defined by one measurement technique, one property and one product [6]. Some areas may contain more than one measurement technique, property or product as long as equivalence and comparability can be demonstrated.

The following aspects should be taken into consideration when a laboratory is establishing its strategy of participation in proficiency testing:

a) 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 participation should be made dependent on the extent to which other measures have been taken. Other types of QA include, but are not limited to:

   - regular use of (certified) reference materials;
   - comparison of analysis by independent techniques;
   - participation in method development/validation and/or reference material characterisation studies;
   - use of internal quality control measures;
   - other inter/intra – laboratory comparisons e.g. analysis on blind samples within the laboratory.

b) The level of risk presented by the laboratory, the sector in which they operate or the methodology they are 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 traceability (e.g. availability of reference materials, national standards, etc...);
   - known stability/instability of the measurement technique;
   - significance and final use of measurement data (e.g. forensic science represents an area requiring a high level of assurance).

c) Different types of PT that can be used by laboratories and should be accepted by accreditation bodies, regulatory bodies or customers, include:

   - PT organised by independent organisations such as accreditation bodies or organisations such as ILAC (International Laboratory Accreditation Cooperation), EA (European co-operation for Accreditation), APLAC (Asia Pacific Laboratory Accreditation Cooperation) and IRMM (Institute for Reference Materials and Measurements);
   - ILC organised by a sufficient number of laboratories (none of which is independent) as a single or continuous exercise;
   - submission of an internal sample or object to another or more external laboratories for the purposes of data comparison.

d) 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. For some fields PT may only be possible or economically feasible for parts of the test/calibration undertaken (e.g. EMC tests on simple objects for a limited number of quantities to be measured). In these areas the suitability of other QA/QC measures is paramount.

e) Any legislative requirements for frequency of type of participation.

When determining an area of technical competence (sub-discipline) it may be helpful to consider a stepwise approach working up from measurement technique through properties to products. This is because it is more likely that there will be several products and/or properties associated with one measurement technique within a given sub-discipline than vice versa:

a) with reference to the measurement technique: it is possible but not common to include different measurement techniques in the same sub-discipline;

b) with reference to the property to be measured, determined or identified: it may be possible to include more than one property (parameter) in the same sub-discipline;

c) with reference to products to be measured: It may be possible to include different products in the same sub-discipline provided that the matrices, objects or materials included, are of equivalent nature.

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

5.2 Availability of PT schemes

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

a) The EPTIS database (www.eptis.bam.de) lists hundreds of PT schemes operated around the world. The focus is on the field of testing;

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

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

d) The PT providers in the participant's own country, which probably will also have (summarised) information about the PT schemes of other providers;

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

5.3 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, one should take into account that different PT schemes will provide different levels of fitness for purpose and that it is rare that a PT scheme with a perfect fit for a laboratory 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 is not limited to the following questions:

a) Are the number of samples, matrices and/or levels offered by the PT scheme similar to those encountered in the laboratory?

b) Is the statistical design described and does it take the different measurement methods into consideration?

NOTE 1 Statistical design covers the process of planning, collection, analysis 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 by.

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

d) Is the frequency of rounds sufficient?

e) Does the PT providers comply with the requirements of ISO/IEC 17043 [4]?

NOTE Accreditation of a PT provider by a National Accreditation Body provides additional confidence of their competence.

f) PT schemes have also an important educational role. If the PT provider treat the data on the basis of different analytical methods and treat also the uncertainty results reported by the participants, this can give valuable information for the further evaluation of analytical methods.

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. A simple form, as given in appendix A, can be helpful to systematically perform and record, this selection process.

If the PT scheme and the needs of the laboratory are sufficiently comparable, the laboratory should then seriously consider participating. 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 also is a good option to judge the relevancy of a specific PT scheme.

If the laboratory can choose its frequency of participation, since some PT providers offer a flexible participation, the stability and the criticality of the method should be taken into consideration, along with the throughput of routine measurement samples in the laboratory.

6 USE OF PT BY LABORATORIES

The basic use of PT for a laboratory is to assess its performance for the conduct of specific measurements or calibrations.

The results and information received from the participation in PT schemes will provide laboratories with either a confirmation that the laboratory's performance is satisfactory or an indication that there are potential problems and that corrections should be made.

However, the use of PT should be much wider than the basic statement of whether the laboratory is competent or not. The laboratories can, as mentioned in 4.1, benefit from the participation in PT schemes in many ways [8]:

1. Identifying Measurement Problems (as a risk management and performance improvement tool)

If a laboratory's result in a PT scheme indicates unsatisfactory performance t, this should start a process of investigation of potential sources of error. 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 clients 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 clients 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.

2. Comparing Methods or Procedures

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

The PT scheme itself might, in some cases, provide summaries and comparisons of all laboratories' methods or commercial kits, for example like for the EU vigilance procedure related to in-vitro diagnostic medical devices [9]. For new or unusual activities, such data could be most valuable and assist the future selection of appropriate methodology by the laboratory or indicate the need for additional investigation before adoption of new methods.

3. Comparing Operator Capabilities

When sufficient PT items are available to more than one operator within a laboratory, the laboratory has the added benefit of being able to compare the results of its operators. This can assist the laboratory to not only compare the performance of its own operators, but might also provide some inputs to the laboratory's estimates of its measurement uncertainty for the relevant measurements.

This might also allow the laboratory to evaluate the between-operator repeatability achieved by the laboratory compared to published (or otherwise available) data for the measurements concerned.

The PT scheme itself might, in some cases, enable results to be reported by more than one operator.

4. *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 organisation.

5. 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 measurement could be improved. This might, for example, include additional operator training, adoption of new or modified methods, enhancing internal quality control of data, equipment modifications, calibration or replacement etc.

6. Educating Staff

Many PT schemes have, as one of their objectives, provision of information on methodology, 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.

7. 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.

8. 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 their laboratory's staff successful performance, often in areas of critical significance to their organisation's activities and responsibilities.

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

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 organisations.

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 or not the laboratory needs to participate for accreditation purposes.

9. 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 or PT rounds.
The "PT approach" can, in specific cases, also be used to evaluate the uncertainty. For example, if the same method is used by all the participants to 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 a collaborative study [10], [11] .

In appendix H, there is an example of how the laboratory's measurement uncertainty is verified through the participation in a PT scheme.

10. Use of PT items as Internal Quality Controls

In some PT schemes, where there is sufficient, stable material provided to participants, the un-used material could be useful for internal quality control monitoring of measurements as a form of reference material.

Where appropriate, the reference values assigned to the PT item (or the consensus values achieved during the PT scheme) might be considered useful as internal reference values for quality control of measurement, operator training, etc.

11. Verification of Method Performance

Depending on the design, some PT schemes will be useful in determining the precision (repeatability and reproducibility) or comparative trueness of the methods used in the PT scheme. In most cases, determination of precision and trueness of the methods is not the primary aim of the PT scheme. To this, further information is often needed and may be obtained from the PT provider.

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 the PT providers, so that the laboratories can better understand the evaluations performed. This should help the laboratory in the selection of the appropriate 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 [12], 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 be contaminated with 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 from the data set. However, the more common approach now used by PT providers, as recommended in ISO 13528 [12], is to use robust statistics [13, 14]. Robust statistics has the advantage of reducing the contribution of outliers to the calculated statistical parameters such as the mean and standard deviation. There a number of robust statistical approaches, some of which are described in ISO 13528 [12].

7.2 Basic elements for the evaluation of PT results

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

1. The assigned value.

2. The standard deviation for proficiency assessment.

In addition the PT provider would be expected to provide an estimate of the measurement uncertainty and a statement of the metrological traceability of the assigned value, as this concept has been included in ISO/IEC 17043 [4]. The relevance, need and feasibility of this estimation shall be determined by the design of the PT scheme.

Different methods can be used to establish these values [12, 15]. 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 the reference documents. 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 when selecting a PT scheme.

7.2.1 Assigned value

There are, as described in ISO 13528 [12], essentially five methods available to obtain the assigned value, a working estimate of the true value:

1. Formulation.
2. Certified reference values.
3. Reference values.
4. Consensus values from expert laboratories.
5. Consensus value from participants.

The description of the assigned value is given in Appendix D.

7.2.2 Standard deviation for proficiency assessment

There are, as described in ISO 13528 [12], essentially five approaches to determine the standard deviation for proficiency assessment, i.e. the acceptable range of participant results:

1. Prescribed value.
2. By perception.
3. From a general model.
4. From the results of a precision experiment.
5. From data obtained in round of a PT scheme.

The description of the standard deviation for proficiency assessment is given in Appendix D.

A common way, at present, to establish the assigned value and the standard deviation for proficiency assessment, is the use of the participants PT results to calculate both values. However, it is strongly recommended in the Harmonised Protocol for the PT in Analytical Chemistry issued by the IUPAC [16], that the scoring methods should be based on fitness for purpose criterion, envisaged by the PT provider in the specific application according to the particular circumstances of the determination. Thus wherever possible, the PT provider should base the standard deviation for proficiency assessment 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.

7.2.3 Performance evaluation

Performance evaluation (or score) by the PT provider adds value to the raw analytical results produced 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 types of measurement uncertainty can be taken into account:

1. Measurement uncertainty of the assigned value.
2. Measurement uncertainty of the participant result.

Given the diverse purposes of PT schemes it is not possible to define a single universal evaluation method. Therefore, a number of statistical designs used for the evaluation of performance are available. The most common are listed below and also given in Appendix (E). Other statistical designs, not covered in this document, are given in ISO 13528 [12].

a) "z score" (most commonly used and measurement uncertainty not taken into account);

$$z = \frac{(x - X)}{\hat{\sigma}}$$

where:
x = result reported by participant
X = assigned value
$\hat{\sigma}$ = standard deviation for proficiency assessment

b) "z'-score" (standard uncertainty of the assigned value is taken into account);

$$z' = (x - X) \Big/ \sqrt{\hat{\sigma}^2 + u_X^2}$$

where:
x = result reported by participant
X = assigned value
$\hat{\sigma}$ = standard deviation for proficiency assessment
u_X = the standard uncertainty of the assigned value X

c) "zeta-score" (standard uncertainty of the assigned value and the participants result is taken into account);

$$\zeta = \frac{x - X}{\sqrt{u_x^2 + u_X^2}}$$

where:
u_x = the participant's own estimate of the standard uncertainty of its result x
u_X = the standard uncertainty of the assigned value X

d) "E_n Number" (expanded uncertainty of the assigned value and the participants result is taken into account)

$$E_n = \frac{x - X}{\sqrt{U_x^2 + U_{ref}^2}}$$

where:
U_x = the expanded uncertainty of the participant's result x
U_{ref} = the expanded uncertainty of the assigned value X determined in a reference laboratory

The following judgment is commonly made for z, z' and zeta scores:

a) $|z| \leq 2{,}0$ the score indicates "satisfactory" performance and generates no signal.

b) 2.0 < $|z|$ < 3.0 the score indicates “questionable” performance and generates a warning signal.

c) $|z| \geq$ 3.0 the score indicates “unsatisfactory” performance and generates an action signal.

The following judgment is commonly made for E_n Numbers:

a) $|E_n| \leq 1.0$ the score indicates "satisfactory" performance and generates no signal.

b) $|E_n| > 1.0$ the score indicates "unsatisfactory" performance and generates an action signal.

The basis of the evaluation must be consistent from round to round of a PT scheme, so that scores in successive rounds are comparable. Only in this way can a participant see long-term trends in his performance.

7.2.4 Effect of the uncertainty of the assigned value

The standard uncertainty of the assigned value depends on the method that is used to derive it, and also, when it is derived from measurements in several laboratories, on the number of laboratories and, perhaps, on other factors. Methods for calculating the standard uncertainty of the assigned value can be found in ISO 13528 [12].

If the standard uncertainty (u_X) of the assigned value is too large in comparison with the standard deviation for proficiency assessment, 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 $u_X \leq 0{,}3\,\hat{\sigma}$ [16], 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 should have taken one of the following steps:

a) used a different method for determining the assigned value such that its 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, zeta-score or E_n number);

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

7.2.5 Qualitative PT schemes

For these PT schemes there is no commonly accepted statistical evaluation, unless the PT provider establishes performance scores by comparison of the laboratories results to the assigned value by transforming the qualitative results into quantifiable data based on predetermined criteria.

If no performance scores are established, the results will mainly be in the form of “yes/no” or “detected/non detected” result. For this type of results, there is not, at present, a recognised common approach.

7.2.6 Outliers

An outlier is an observation that is numerically distant from the rest of the data. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has a high spread and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modelled by a mixture model.

The PT provider should mention how it has taken into account any outliers in the statistical analysis.

8 LABORATORY INTERPRETATION OF PT RESULTS

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 to be evaluated.

8.1 Performance evaluation by the laboratory

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 top management may not always be familiar with PT performance, and it is highly advisable that they gain an appropriate level of understanding of PT's.

As the laboratory should be using validated methods along with internal quality controls, any poor performance is to be taken seriously as it indicates that there is a problem with the validation and/or the internal quality controls.

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 round for a laboratory, where all participants have a satisfactory performance, does not necessarily indicate a high level of competence, the standard deviation for proficiency assessment could be in this case too large. Neither, on the other hand, does one unsatisfactory performance in one 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 laboratories measurement process 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.

8.1.1 Review of single PT round results

The results of each PT round are to be specifically 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 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 process.

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 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 more fit for purpose standard deviation for proficiency assessment (see Appendix F.2).

If after a thorough investigation, the laboratory concludes that the result is indeed unsatisfactory, then corrective actions should be initiated (see chapter 8.2).

The laboratory's results from its participation in PT can also be used to check the validity of the laboratory's measurement uncertainty.

8.1.2 Monitoring PT performance over time

Following the careful evaluation of single round results, the monitoring of PT performance over time should be done, in order to identify potential problems related to imprecision, systematic error or human error.

A graphical plot of performance scores from round to round in order to monitor PT 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 internal quality control (IQC) procedures would normally be expected to identify trends associated with, for example, improper instrumental calibration or maintenance, or use of reagents. Monitoring PT performance over time acts as a complementary system for this.

To decide whether the performance is improving or decreasing with time, the data from subsequent rounds has to be comparable. However, the data set from the same measurement from two different PT rounds may have a different standard deviation for proficiency assessment and so lead to the performance scores being calculated differently, depending on the group of laboratories that participated and the influence of other variables arising from difficult or complex samples. The participants may calculate their own z-scores (or other performance scores) using a selected standard deviation for proficiency assessment, if the value used in consecutive rounds differs substantially. A standard deviation for proficiency assessment from literature (e.g. from a standard method like ISO, EN, ASTM, DIN) can be used. If such values from literature are not available, the laboratory's own criteria may be chosen depending on the goal of the participation to the PT or the importance of the measurement (any realistic value can be used, e.g. 10% of the assigned value). Note that the selected standard deviation for proficiency assessment does not have to be constant, i.e. it may be concentration dependent. If a laboratory decides to recalculate their own performance scores, they should justify and document their choice.

Examples of how an individual laboratory can monitor its performance over time is given in Appendix G.

8.2 Investigation of unsatisfactory or questionable PT results

8.2.1 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, which can include, the criticality of the method, 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.

For questionable results, when participating in a continuous PT scheme, with several rounds/year, the following criteria can, for example, be chosen by the laboratory:

a) 2 consecutive questionable performance scores for the same parameter;

b) 9 consecutive performance scores, for the same parameter, which have the same bias sign against the assigned value.

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

8.2.2 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. An example of a form supporting this approach is given in Appendix B:

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

a) analyse the problem based on the raw data, the overall performance of the round, the result of successive interlaboratory studies and internal quality control 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.

8.2.3 Causes for poor performance

The reasons for obtaining a poor performance are "unfortunately" numerous, 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. The following main causes, in order of importance, were found during a web-based survey [17]:

a) sample preparation (e.g. weighing, drying, extraction, digestion, clean-up, dilution, etc);
b) equipment failure or servicing problem;
c) human error (e.g. inappropriate training, transcription);
d) calibration;
e) selection of measurement method;
f) calculation error;
g) reporting problem (e.g. format, units, detection, interpretation);
h) PT item problem;
i) sample transport and storage;
j) primary sampling;
k) other problem category;
l) sample tracking (e.g. labelling, chain of custody);
m) PT scheme provider problem.

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

a) clerical error;
b) technical problem (e.g. method, equipment, training, internal quality controls);
c) problem related to the PT scheme (e.g. inadequate scheme, inappropriate evaluation).

It may be possible that after a thorough investigation, the origin of the poor performance is not identified. A single poor performance could then be attributed to random error. If it is a repeated poor performance, then the analytical process should be questioned.

8.2.3.1 Clerical error

Although clerical errors are not directly linked to the laboratory's technical competence, it 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) incorrect units;
d) decimal error.

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.

8.2.3.2 Technical problem

Due to the complexity of analysis, problems can occur at every level of the analytical procedure and each of the following steps of the analytical process should be reviewed during the investigation:

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

b) method/internal quality control data;

c) equipment/reagents/calibration;

d) environmental conditions;

e) data processing.

If the investigation of the analytical process does not enable the laboratory to identify the root cause, it may be necessary to review the method validation.

8.2.3.3 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) parameter concentration levels outside the scope of application of the method;

d) lack of stability or homogeneity of the samples;

e) inappropriate instructions to participants;

f) PT item storage problems;

g) inappropriate peer group;

h) inappropriate assigned value;

i) inappropriate standard deviation for proficiency assessment;

j) incorrect data entry from 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.

Appendix A: Selection form for a relevant PT scheme

Parameter:

PT item:

Method:

PT provider:

PT scheme evaluated:

Selection Criteria	Y	N	NA	Remark
Is it a critical parameter for the laboratory?				
Are the parameters proposed equal or equivalent to the ones routinely tested?				
Is the PT item equal or equivalent to the one routinely tested?				
Is the parameter range appropriate to the laboratory?				
Does the PT provider treat the results obtained by taking into account the different methods used?				
Is the number of participants or the size of my peer group appropriate?				
Is the frequency of rounds sufficient?				
Is it possible to report the result's uncertainty?				
If yes, does the PT provider take into account the uncertainties reported by the participants in its statistical analysis?				
Does the PT provider report the uncertainty of the assigned value?				
Is the metrological traceability of the assigned value given?				
Does the PT provider give information about the statistical design used?				
Is the evaluation of the performance of the laboratories based on a scoring criteria (e.g. z-score)				
Does the PT provider provide assistance in the case of poor results?				
Do the reports include sufficient analysis of results and information for laboratories to carry out corrective actions?				
Does the PT provider provide "surplus/repeat samples" to laboratories for carrying out corrective actions?				
Is the PT provider accredited or recognized by a third party?				
Are the reports edited in a language understood by the laboratory?				

Conclusion:

Is this PT scheme relevant for the laboratory: ☐ yes ☐ no

Remark:

Date: Approved by:

Appendix B: Form for documenting PT investigations

Investigation performed by: Date:

Parameter:
PT item:
Method:
PT scheme: Round N°:
Laboratory's result(s): Acceptable result/ range:
Performance evaluation (score):
Parameter critical: □ YES □ NO
How relevant is the PT scheme compared to routine analysis (e.g. matrix, parameters, concentration level, etc)?

Do the results of previous rounds in the PT scheme indicate a questionable or unsatisfactory trend? If yes, analysis of this trend should be provided:

Initial evaluation:

Was the PT item received in a satisfactory condition?	□ YES	□ NO
If no, could this condition explain the poor result?		
Was the PT item equivalent to a routine sample?	□ YES	□ NO
If no, could this explain the poor result?		
Was the PT item tested as a routine sample?	□ YES	□ NO
If no, could this be the cause for the poor result?		
Is the evaluation based on results grouped according to method?	□ YES	□ NO
If yes, can this explain the poor result?		
Based on the comments given above, should the relevancy of the PT scheme be reviewed?	□ YES	□ NO
Was the initial PT item remeasured after receipt of PT evaluation?	□ YES	□ NO
If yes, is the result comparable?		
Was a repeat item requested and remeasured?	□ YES	□ NO
If yes, is the result comparable?		

Clerical Investigation:

Typical clerical errors can be for example: transcription errors, data entry error, PT provider not informed of method change, incorrect units used.

Was the poor result due to a clerical error?	□ YES	□ NO
Corrective action taken, if any?		
Corrected result:		
Is this result still questionable or unsatisfactory?	□ YES	□ NO
If yes, the investigation should be continued.		

Technical Investigation:

The following aspects should be taken into consideration:

Analytical procedure:
Internal quality controls:
Storage/preparation of the PT item:
Equipment:
Environmental conditions:
What impact is there on past and future routine results?
Conclusion:

Corrective action(s) taken:

Approved by :	
Technical Manager:	Date:
Quality Manager:	Date:

Appendix 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 quality control activities are one 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 appropriate PT schemes in order to have an independent evaluation of the performance of the laboratory.

It is the responsibility of the laboratory to ensure that when providing their 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. Scheme protocols and other documentation will be studied and, if necessary, the PT provider contacted to discuss or clarify any outstanding issues. The level of performance on 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 satisfy themselves 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, that scheme 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 scheme for their own purposes, it is recommended that they discuss fully the scope and operational parameters of the 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 analytes in all 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 determinations in which they have an interest are clearly stated, as the scheme may have a broader scope, and performance of laboratories in determinations 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 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 round of a PT scheme does not make a laboratory poor, neither does the achievement of 100% satisfactory results in any 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.

Appendix D: Statistical aspects for PT

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, often a consensus value.

2. The standard deviation for proficiency assessment

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

D.1 Assigned value and standard uncertainty of the assigned value

There are, as described in ISO 13528 [12], essentially five methods available to obtain the assigned value and its associated standard uncertainty:

1. Formulation: the addition of a known amount or concentration of 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 include, inter alia, extraction and speciation] where recovery problems may well arise.

The standard uncertainty is estimated by combination of uncertainties using the approach described in the Guide to the expression of Uncertainty in Measurement (GUM) [18].

2. Certified reference values: when the PT item is a certified reference material (CRM), its certified reference 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.

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

3. Reference values: a selection of the prepared PT items is measured, by a chosen laboratory, either using a primary method or alongside a certified reference material (CRM). The assigned value is derived directly from the primary method used or from a calibration against the certified reference value of the CRM. This will provide a traceable 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 standard uncertainty is derived from the test results of the chosen laboratory, and the uncertainties of the certified reference values of the CRM.

4. Consensus values from expert laboratories: the determination of a consensus value obtained from the outcome of a group of expert or referee laboratories being proficient in the analytical methods applied. This is probably the closest approach to obtaining true values for the PT items, but it may well be expensive to do so. Another problem is that it is often hard or even impossible to find a group of expert or reference laboratories whose expertise is beyond doubt and accepted by all participants of the PT. This is even more true for large, international PTs with participants from many countries. For a number of analyses, the true value is, in principle, defined by the method used. In these cases, the expert or referee laboratory should all use the same method 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 methods applied should be declared before the PT is set up.

If each expert laboratory reports an estimate of the standard uncertainty, then the standard uncertainty is estimated by:

$$u_X = \frac{1.25}{p} \times \sqrt{\sum_{i=1}^{p} u_i^2}$$

where:
u_i = the standard uncertainty from the expert laboratory
p = number of expert laboratories

If not the standard uncertainty is estimated as, below, for the consensus value from participants.

5. Consensus value from participants: the use of a consensus value, produced in each round of the PT, 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 samples, is often the only way to establish an estimate of the true value.

The standard uncertainty is estimated as:

$$u_X = 1.25 \times s^* / \sqrt{p}$$

where:
s^* = robust standard deviation of the participants results.
p = number of participants

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 methodology and this bias will not be reflected in the standard uncertainty of the assigned value calculated as described above.

D.2 Standard deviation for proficiency assessment

There are, as described in ISO 13528 [12], essentially five approaches to determine the standard deviation for proficiency assessment i.e. the acceptable range of participant results:

1. Prescribed value: the standard deviation for proficiency assessment may be set at a value required for a specific task of data interpretation, or it may be derived from a requirement given in legislation.

With this approach, the standard deviation for proficiency assessment becomes equivalent to a “fitness for purpose” statement for the measurement method.

2. By perception: the standard deviation for proficiency assessment may be set at a value that corresponds to the level of performance that the PT provider would wish laboratories to be able to achieve.

With this approach, the standard deviation for proficiency assessment becomes equivalent to a “fitness for purpose” statement for the measurement method.

3. From a general model: the value of the standard deviation for proficiency assessment may be derived from a general model for the reproducibility of the measurement method.

A disadvantage of this approach is that the true reproducibility of a particular measurement method 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 parameter, and not on the parameter, the measurement procedure, or the sample size.

4. From the results of a precision experiment: when the measurement method to be used in the PT scheme is standardized, and information on the repeatability and reproducibility of the method is available, the standard deviation for proficiency assessment may be calculated using this information.

5. From data obtained in round of a PT scheme: with this approach, the standard deviation for proficiency assessment used in a round of a scheme is derived from the results reported by the participants in the same 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 round to round, making it difficult to use values of the z-score for a laboratory to look for trends that persist over several rounds.

Appendix E: Performance evaluation for PT

The most common statistics, already highlighted in 7.2.3, used for the evaluation of performance are given below. Other statistical designs, not covered in this document, are given in ISO 13528 [12].

E.1 "z-score"

One of the basic and common elements in all PTs is the use of a performance indicator to quantify the analytical performance of each participant [12, 16]. The z-score is frequently advised as such a performance indicator. The z-score is a measure of the deviation of the result from the assigned value and is calculated as:

$$z = \frac{(x - X)}{\hat{\sigma}}$$

where:
x = result reported by participant
X = assigned value
$\hat{\sigma}$ = standard deviation for proficiency assessment

The main assumption in using the z-score is that the individual z scores will have a Gaussian or normal distribution with a mean of zero and a standard deviation of one. On this basis analytical results can be described as 'well-behaved'. A common classification based on z-scores can be made.

E.2 "z'-score"

$$z' = (x - X) \Big/ \sqrt{\hat{\sigma}^2 + u_X^2}$$

where:
x = result reported by participant
X = assigned value
$\hat{\sigma}$ = standard deviation for proficiency assessment
u_X = the standard uncertainty of the assigned value X

When deciding whether to use either z-scores or z′-scores, the PT provider shall consider the following aspects:

a) if $u_X \leq 0{,}3\,\hat{\sigma}$ [16], then the standard uncertainty of the assigned value is negligible and it is unlikely that there will be any benefit from the use of z′-scores;

b) when $u_X \geq 0{,}3\,\hat{\sigma}$ the standard uncertainty of the assigned value is not negligible and it is recommended to use z′-scores;

c) how severe are the consequences to laboratories when their results give rise to warning or action signals? Are the results used to disqualify laboratories from carrying out the measurement method for some group of users?

E.3 "Zeta-score"

Increasingly, laboratories are being encouraged to estimate the uncertainty of their results and as such it is becoming more common for PT providers to incorporate such information into PT performance scoring.

$$\zeta = \frac{x - X}{\sqrt{u_x^2 + u_X^2}}$$

where:
u_x = the laboratory's own estimate of the standard uncertainty of its result x
u_X = the standard uncertainty of the assigned value X

The following performance criteria are commonly recognized for z, z' and zeta scores:

a) $|z| \leq 2{,}0$ the score indicates "satisfactory" performance and generates no signal.

b) $2.0 < |z| < 3.0$ the score indicates "questionable" performance and generates a warning signal.

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

E.4 "E_n NUMBER"

Another alternative scoring system is the use of an "E_n" number, which takes into account the expanded uncertainty:

$$E_n = \frac{x - X}{\sqrt{U_x^2 + U_{ref}^2}}$$

where:
U_x = the expanded uncertainty of the participant's result x
U_{ref} = the expanded uncertainty of the assigned value X determined in a reference laboratory

The E_n number is used when the assigned value has been produced by a reference laboratory. However, it is necessary for participants to have a good understanding of their uncertainty and for each participant to report it in a consistent way.

The following judgment is commonly made for E_n numbers:

a) $|E_n| \leq 1.0$ the score indicates "satisfactory" performance and generates no signal;

b) $|E_n| > 1.0$ the score indicates "unsatisfactory" performance and generates an action signal.

The critical value of 1, rather than 2 used for z-scores, is used because the E_n number is calculated using expanded uncertainties instead of standard uncertainties.

E.5 Scoring when there is no common fitness for purpose criterion – modified z-scores:

Occasionally participants in a PT scheme find that the standard deviation for proficiency assessment used by the PT provider is inappropriate for some or all of the applications that they are engaged in. In consequence the resulting z-score would be misleading. In such instances the participant can validly construct an individual scoring system based on an appropriate standard deviation for proficiency assessment that is fit for an individual purpose.

That is [16]:

$$z = \frac{(x - x_{av})}{\sigma_f}$$

where
x_{av} = the assigned value
σ_f = the appropriate value of standard deviation for proficiency assessment

This does not, however, require action from the PT provider.

Appendix F: Examples of different performance evaluation approaches

F.1 Comparative results of assigned value and interlaboratory precision between classical vs. robust statistical methods.

In the development of this comparison, the statistical protocol was conducted to achieve the estimation of both the true value and the standard deviation for proficiency assessment. According to that, a couple of statistical approaches had been considered in order to evaluate the analytical results received from the participants (Table F.1): one conventional method based on ISO 5725 [1] calculations and a second approach based on the application of robust statistics by means of the algorithm A as described in ISO 13528 [12].

With regard to the assigned value, no relevant differences (less than 0,2 ‰,) were found between the statistical methods applied, so the robust average of all the participant results without outlier detection was chosen as the best estimation of the true value of the concentration of analyte of interest in this scheme (robust average = 586.7 ‰).

However, as for the estimation of the interlaboratory precision, the two above-mentioned protocols were compared, leading to some relevant differences in terms of the standard deviation for proficiency assessment. As a result of that, the statistical treatment of the results revealed that following classical methods the relative standard deviation in percent (%RSD) value was roughly twice as large, than the one obtained according to robust estimation with the same consideration about outliers that were expressed previously.

Table F.1: Comparative results of assigned value and interlaboratory precision between classical vs. robust statistical methods.

	ISO 5725	ISO 13528
Assigned value	586.5 ‰	586.7 ‰
Standard deviation for proficiency assessment	0.56 ‰	0.31 ‰
RSD	0.10	0.05

Some robust statistical procedures choose the median and normalized interquile range IQR, which are measures of the centre and spread of the data (respectively), and are used similarly to the mean and the standard deviation. As robust statistics, the median and normalized IQR are less influenced by the presence of outliers in the data.

The median is the middle value of the group, i.e. half of the results are higher than it and half are lower. If N (total number of results) is an odd number, the median is the single central value. If it is even it is the mean of the two central vales.

The normalized IQR is a measure of variability of the results and is equal to the IQR multiplied by a factor (0.7413), which makes it comparable to a standard deviation.

F.2 Discussion on the assessment of laboratories performance.

This topic is intended to explain a wider range of possibilities that the PT-scheme provider might consider to evaluate the data submitted by the participants, so a number of cases with a different approach were estimated to cover these statistical principles [19], [20], [21].

On the whole, in order to evaluate the assessment of each laboratory performance by interpreting z-score values, the values considered in the expression of z-score (assigned value and standard deviation for proficiency assessment) have been calculated as follows in each one of the four statistical approaches:

1. ISO 5725 [1]: general mean and reproducibility standard deviation, with outlier detection;

2. Median and NIQR method: median of the whole data and normalized interquartile range [20];

3. ISO 13528: robust average and robust standard deviation calculated according to algorithms A and S, without outlier detection [12];

4. Fit-for-purpose criterion: robust average and a target standard deviation for proficiency assessment value according to a fixed %RSD from appropriate past PT-rounds at this level of concentration [14].

After calculation by each one of the four protocols considered (Table F.2), it can be stated that fourteen participants show z-score values considered as acceptable ($|z| \leq 2$) regardless of the statistical method applied. The reason for this behaviour lies in the fact that they are the laboratories that provide the most balanced results with fewer deviations in data spread due to common analytical techniques, so the statistical protocol has not influenced their performance.

Furthermore, in terms of distribution of z-score values corresponding to the other five laboratories, a certain trend in the spread is revealed. Thus, according to ISO 5725 [1], z-score values comply with the acceptance criteria, particularly because the data from one laboratory has been rejected as outliers.

Table F.2: Summary of the overall z-score results obtained by participant laboratories reported following the different statistical protocols.

Participant	ISO 5725	Median & NIQR	ISO 13528	Fit-for-purpose
Lab.31	outlier	-4.76	-8.27	-5.68
Lab.06	-1.84	-2.28	-3.86	-2.65
Lab.14	-1.63	-2.06	-3.48	-2.39
Lab.08	-1.52	-1.95	-3.28	-2.25
Lab.13	-1.30	-1.73	-2.89	-1.98
Lab.27	-0.09	-0.50	-0.69	-0.48
Lab.22	0.15	-0.26	-0.27	-0.19
Lab.05	0.16	-0.25	-0.26	-0.18
Lab.03	0.34	-0.06	0.08	0.06
Lab.20	0.40	0.00	0.19	0.13
Lab.15	0.53	0.13	0.43	0.29
Lab.21	0.56	0.16	0.47	0.32
Lab.29	0.56	0.16	0.47	0.32
Lab.12	0.62	0.22	0.58	0.40
Lab.10	0.65	0.25	0.64	0.44
Lab.07	0.74	0.34	0.80	0.55
Lab.16	0.81	0.41	0.92	0.63
Lab.24	0.86	0.46	1.01	0.70
Lab.30	1.10	0.71	1.45	1.00

On the other hand, in order to avoid the influence of extreme results, the application of robust statistical methods [22], [13] brings about significantly larger z-score values since no outlier elimination is applied. Accordingly, when the calculation is performed by using the median & NIQR method, z-score values are slightly smaller than the corresponding ones estimated following the robust method based on ISO 13528 [12]. In this case, one laboratory is given a warning signal whereas four laboratories show z-score values considered to give action signals, so that special investigation is required for the laboratory previously considered as an outlier in the parametric approach.

Lastly, when applying a fit-for-purpose criterion [14] according to an end-user requirement where the performance ratio is determined by the PT-provider itself and no outlier rejection of data has been considered due to the own statistic protocol formulation, it can be seen that z-score values give results considered as satisfactory for fifteen laboratories, results considered as warning signals for three participants and that one single laboratory would be allocated an action signal that requires further investigation.

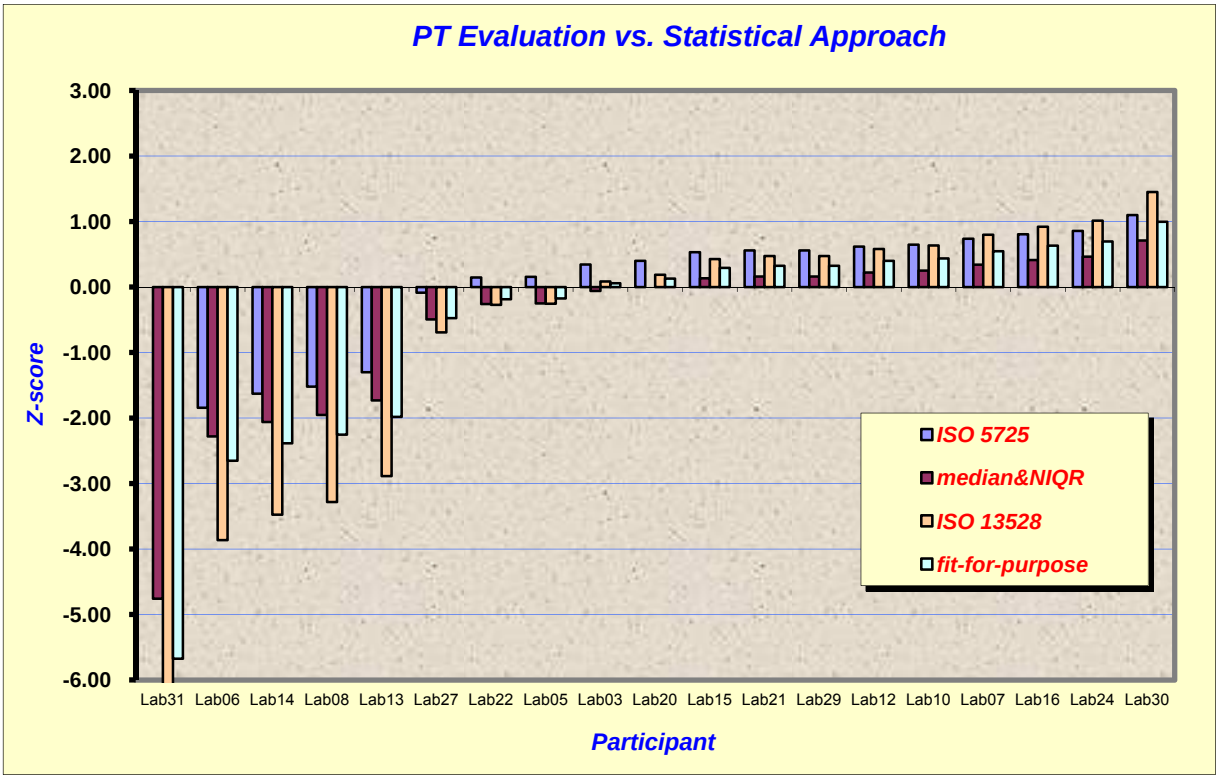


Figure F.1: Graphic summary of the overall z-score results obtained by participant laboratories reported following the different statistical protocols.

F.2.1 Concluding remarks

Due to the fact that PT participant data is usually heavily-skewed and the presence of outliers is very common, classical statistics can lead to overestimations of the standard deviation for proficiency assessment in which many z-scores are considered as satisfactory, unless robust statistical methods are applied, since they are more insensitive to anomalies. This is clearly illustrated in the comparative statistics provided on example data sets given above.

The robust protocols discussed in this document are particularly applicable to normal distribution data with no more than 10% of outliers and which are unimodal and roughly symmetric, apart from cases where it is assumed that all participants do not have the same analytical performance. However, it is observed that the median & NIQR method is more robust for asymmetric data, while in cases of multimodal or skewed distribution of data, the application of mixture models and kernel density functions should be considered.

In this example, the application of both classical and robust statistical methods when dealing with PT data clearly shows that mean values are quite similar, whereas significant differences in the standard deviation value have been found, in some cases too large for fitting the objective of the PT.

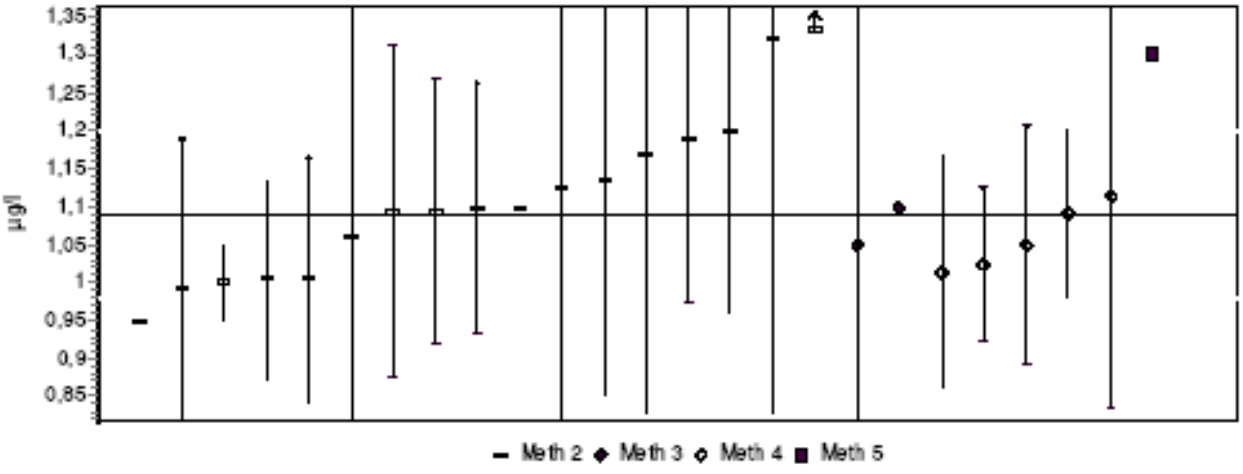
Furthermore, it is quite important to obtain an appropriate estimation of the overall standard deviation parameter that allows the precision of the analytical method but also to provide a performance assessment compatible with the intercomparison requirements.

Finally, the application of a fit-for-purpose criterion should describe the end-user requirement and must be consistent from round to round, so that performance scores in successive rounds might be comparable. The specification of a standard deviation for proficiency assessment in terms of relative standard deviation involves more of a quality goal that the data should meet to reflect fitness for purpose, rather than a simple description of the data results.

F.3 Method dependant assigned value

Before the final decision on establishing the assigned value, the PT provider should also take into account possible differences between the results obtained by different analytical methods. For example in the analysis of low metal content in water, the results can show considerable variation, if several methods have been used. The ICP-MS method is the most sensitive in measuring low metal contents and the mean value of the results obtained by this method might be more appropriate in determining the assigned value. Different pre-treatment procedures can also have an effect on results. For example different digestion acids in the analysis of metals in soil or in sludge can result in different mean values, if the results obtained using different acids in pre-treatment are treated separately.

Example F.1: Determination of Cd (mean value = 1.09 µg/l estimated by calculating) using different analytical methods (Meth 2= GAAS, Meth 3 = ICP-OES, Meth 4 = ICP-MS, Meth 5 = others).



Example F.2: The mean values of the results obtained by using different acids for pre-treatment of a sludge sample in a PT for analysis of metals.

Metal	Pretreatment	Mean (mg/kg)	SD	N
Cr	LN1	23.36	4.36	15
	LO1	34.91	10.69	7
Al	LN1	8918	1200	13
	LO1	16451	3778	8

LN1– *digestion with HNO_3 or with $HNO_3 + H_2O_2$*

LO1 – digestion *HNO_3 + HCl (aqua regia)*

Appendix G: Example of long-term performance evaluation

It is important for a laboratory to know the stability of performance of its measurements over time. This example [24], based on an analysis from the haemostasis field, shows a method to assess the long-term analytical performance of any quantitative analysis, as long as there is sufficient data and that different concentrations are used in the PT scheme.

Individual laboratory test results are compared to the consensus value, determined as the mean value of all the tests results or the mean value of a peer method group. In this model the consensus value can be easily replaced by a target or assigned value.

To evaluate the laboratory's performance, two variables are used:

1. Long-term analytical CV (LCV_a)
2. Long-term total bias

The long-term analytical performance is based on linear regression (based on the least squares method) using the consensus value(x) as the independent and the laboratory value (y) as the dependant variable.

G.1 Long-term Analytical Coefficient of Variation (LCV_a)

The long-term analytical coefficient of variation (LCV_a) is based on the variability of the regression line ($S_{y|x}$) and the mean value of all consensus values ($\overline{X}$). To allow comparison of the LCV_a between laboratories it should be calculated after adjustment for the bias (b). Therefore the LCV_a is now calculated using the formula:

$$LCV_a = \frac{\left(s_{y|x} / b\right)}{\overline{X}} \bullet 100\%$$

G.1 Long-term Total Bias

The long-term total bias (B) can be calculated by the formula:

$$\mathrm{B} = \sqrt{\frac{n-1}{n} \bullet \left(\mathrm{b}-1\right)^2 \bullet \mathrm{s}_x^2 + \left(\overline{Y} - \overline{\mathrm{X}}\right)^2}$$

The long term total bias consists of two components, the proportional bias (PB) and the constant bias (CB).

The proportional bias is caused by the deviation of the slope and depends on the variability of the consensus value.

$$\mathrm{PB} = \sqrt{\frac{n-1}{n} \bullet \left(\mathrm{b}-1\right)^2 \bullet \mathrm{s}_x^2}$$

The constant bias reflects the deviation from the consensus value.

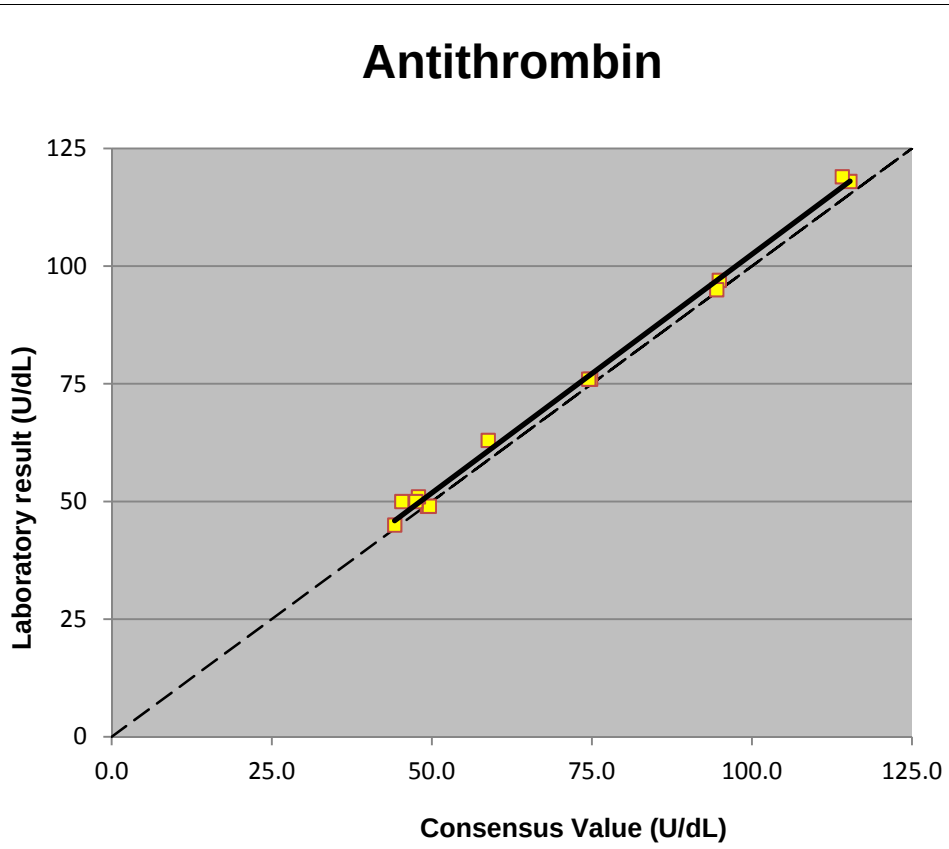
$$CB = \sqrt{(\overline{Y} - \overline{X})^2}$$

The ratio of the proportional and constant bias indicates whether the bias is mainly caused by calibration errors (proportional bias) or by other factors like matrix effects (constant bias).

The long-term evaluation of two parameters is given below:

Table G.1: Long term evaluation of Antithrombin

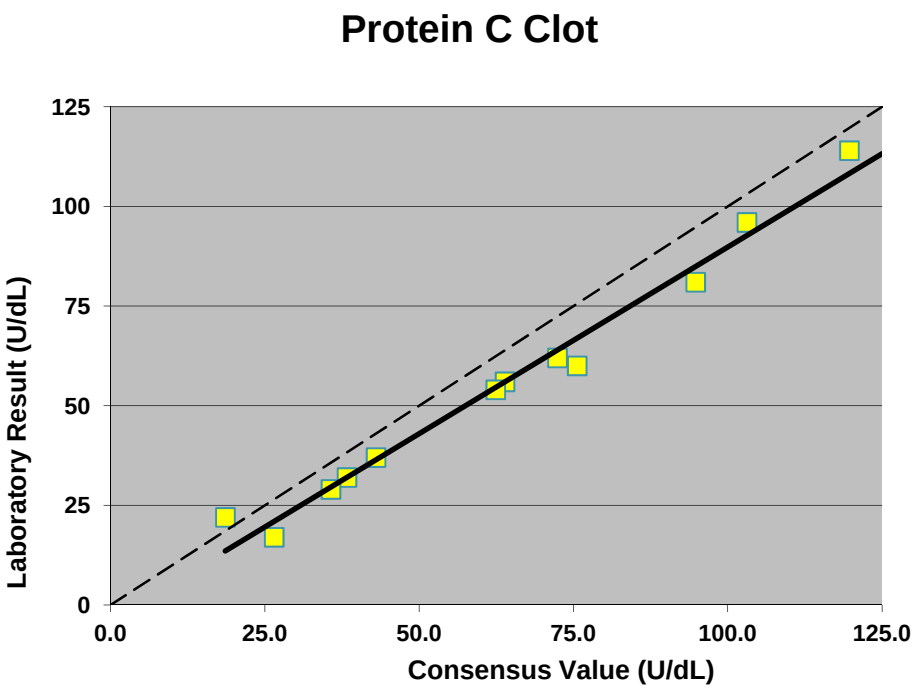
Round	Lab Result (U/dL)	Cons. Value (U/dL)
2005-1		49.8
2005-2	50.0	45.3
2005-3	118.0	115.3
2005-4	49.0	49.3
2006-1	76.0	74.8
2006-2	119.0	114.1
2006-3	49.0	49.6
2006-4	76.0	74.5
2007-1	97.0	94.9
2007-2	51.0	47.9
2007-3	45.0	44.2
2007-4	63.0	58.8
2008-1	50.0	47.5
2008-2	95.0	94.5
Mean	72.2	68.6
SD		26.2
Number	13	



LCVa		2.6%	Intercept	1.08	Variability	1.84
Bias		5.2%	Slope	1.01	Corr.coeff	0.9958

Table G.2: Long term evaluation of Protein C Clot

Round	Lab Result (U/dL)	Cons. Value (U/dL)
2004-1		125.0
2004-2	62.0	72.4
2004-3	22.0	18.6
2004-4	114.0	119.7
2005-1	32.0	38.3
2005-2	60.0	75.6
2005-3	96.0	103.1
2005-4	29.0	35.7
2006-1	81.0	94.8
2006-2	17.0	26.5
2006-3	56.0	63.9
2006-4	37.0	43.0
2007-1	54.0	62.4
Mean	55.0	67.6
SD		35.0
Number	12	



LCVa		7.0%	Intercept	-3.86	Variability	4.46
Bias		18.9%	Slope	0.94	Corr.Coeff	0.9800

Appendix H: Example of the use of measurement uncertainty

In 2008, the Institut für Eignungsprüfung (IfEP) organised a proficiency test “Rockwell hardness test (0801-HRC 2008)”. The participation in this proficiency test was open to accredited and non-accredited laboratories.

This proficiency test was based on the test standard EN ISO 6508-1 [25] and was designed according to ISO/IEC Guide 43-1 [26] and ISO 13528 [12].

The participants received certified reference blocks (certified by MPA NRW Dortmund) on three different levels of hardness (25 HRC, 45 HRC, 60 HRC). Additionally, a piece of standard test material, cold work steel, with a hardness of approximately 50 HRC was send to the participants.

The test procedures were defined as:

Task A:

Make five hardness measurements HRC according to EN ISO 6508-1 [25] on each of the three certified reference blocks.

Task B:

For the evaluation of the measurement uncertainty: Prepare one piece of standard test material and make five hardness measurements according to EN ISO 6508-1 [25].

Participants:

76 laboratories located in 28 countries participated in this proficiency test. 57 participants declared to have an accreditation according to ISO/IEC 17025 [2].

H.1 Evaluation of the proficiency test

The evaluation of the results was based on ISO 13528 [12] and ISO/IEC Guide 43-1 [26]. Task A is evaluated on the criteria “permissible error of the testing machine” and “permissible repeatability of the testing machine” based on EN ISO 6508-2, Table 5 [25]. The results of task B were used for the calculation of measurement uncertainty only. The error of the testing machine E is calculated according to equation (1):

(1) $$E = \overline{H} - \overline{X}_{CRM}$$

$\overline{H}$ is the (arithmetic) mean value of five measurements on a given hardness block.

$\overline{X}_{CRM}$ is the certified reference value of each individual hardness block. The results of this exercise can be seen in figure H.1.

The permissible error of the testing machine GA (1a) is stated in EN ISO 6508-2, table 5 [25]:

(1a) $-1{,}5\ HRC \leq E \leq 1{,}5\ HRC$

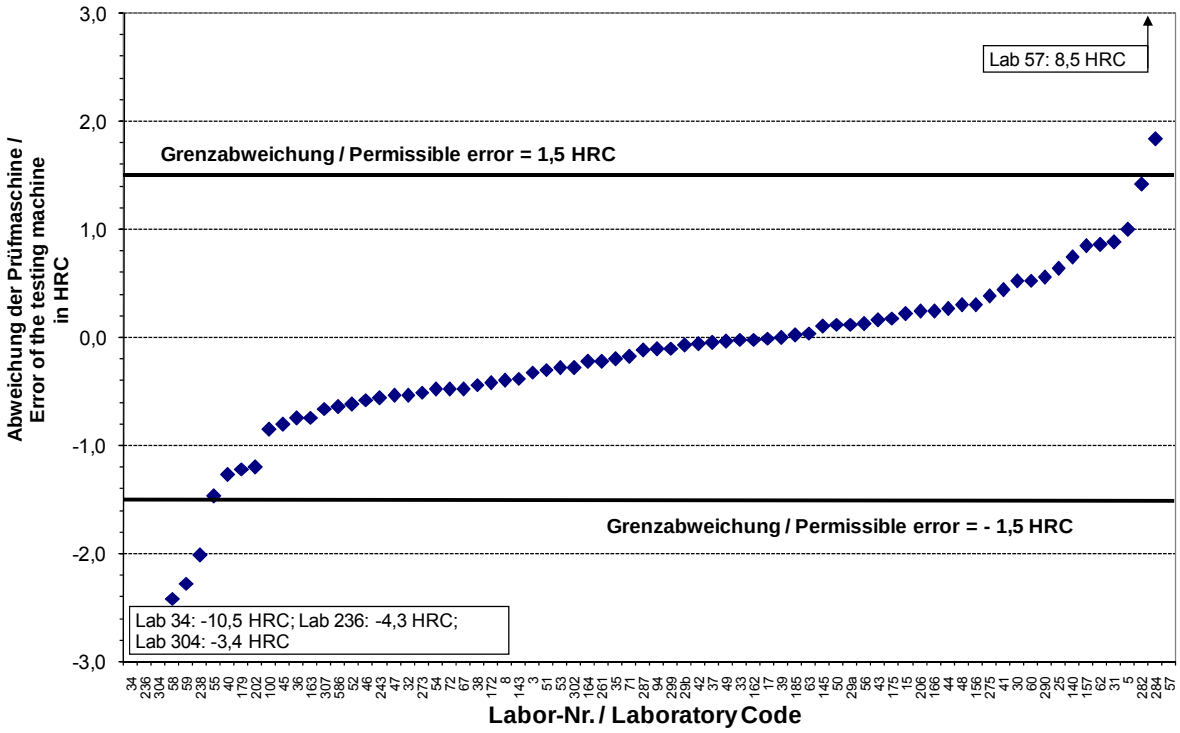


Figure H.1: Test A, Example 45 HRC-Level: Error of the testing machine

H.2 Determination of measurement uncertainty

The evaluation of the measurement uncertainty shall strictly follow the EN ISO 6508-1 [25] method 1 (abbreviated M1) approach which is based on the UNCERT Code of Practice Nr. 14 [28]. Additionally to the test of certified reference material this method requires five measurement of the hardness on a standard material.

All participants were asked to report their expanded measurement uncertainty for the measurement done on the standard material (task B). IfEP calculated for all participants the individual measurement uncertainty according to equations (4) and (5) based on [29] and [30].

$$U = 2 * \sqrt{u_E^2 + u_{CRM}^2 + u_H^2 + u_x^2 + u_{ms}^2} \qquad (4)$$

$$\tilde{U} = \frac{U}{\overline{X}_{CRM}} * 100\% \qquad (5)$$

Where:

U = Expanded measurement uncertainty

$\tilde{U}$ = Relative expanded measurement uncertainty

u_E = Standard measurement uncertainty of the testing machine

u_{CRM} = Standard measurement uncertainty of the certified reference block

$u_{\overline{H}}$ = Standard measurement uncertainty of the laboratory testing machine measuring the hardness of the certified reference block

$u_{\overline{x}}$ = Standard measurement uncertainty resulting from testing the material

u_{ms} = Standard measurement uncertainty according to the resolution of the testing machine

$\overline{X}_{CRM}$ = Certified reference value of the certified reference block

The minimum level of relative expanded measurement uncertainty $\tilde{U}$ is given by the combination of the fixed factors $u_{E,}$ u_{ms} and u_{CRM}. It is at least 2,1 %. The results of measurement uncertainty values reported by the participants can be seen in figure 2. Furthermore, the calculations done by IfEP are shown to demonstrate the differences. For participants, who did not provide all the requested information, the measurement uncertainty was not calculated.

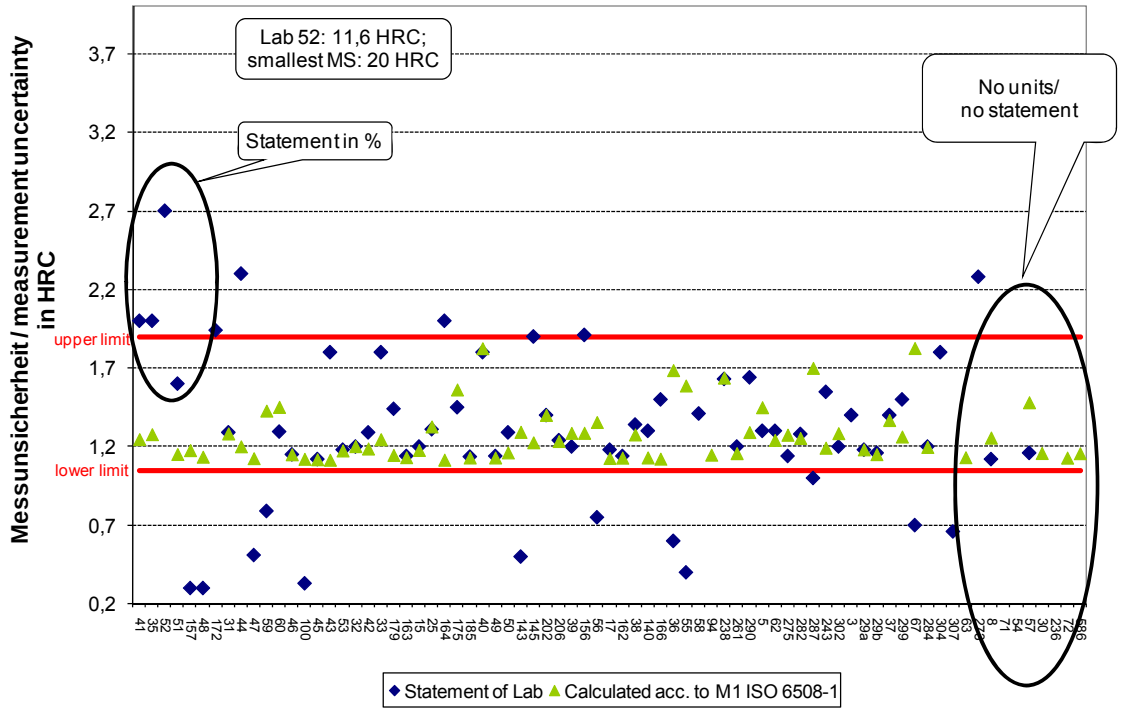


Figure H.2: Comparison of measurement uncertainty stated by participants and calculated by IfEP according to method M1, ISO 6508-1.

H.3 Results and discussion

The measurement uncertainty values reported by participating laboratories in the PT show a good correlation between their results and results calculated by the organiser based on the procedure of the test method. About 20% of all participants reported values which were too small and were not reasonable based on fixed factors which cannot be lower than a defined minimum. These

participants should verify their method of measurement uncertainty estimation and learn from the outcome of this proficiency test. Participants which did not state all the necessary values for the calculation, or overestimated the measurement uncertainty, now have the tools to verify their methods. For this specific test method the calculation of the related measurement uncertainty is based on the test standard. The provider of this proficiency test was able to use the given formula to reflect not only the test result but also the acceptable range of measurement uncertainty values to the participants. The learning effect of such a proficiency test is higher than a PT only reporting performance statistics. In the future, participants in PTs should select the provider based on added value opportunities.

Bibliography

[1] ISO 5725-1994, *Accuracy (trueness and precision) of measurement methods and results*

[2] ISO/IEC 17025:2005, *General requirements for the competence of testing and calibration laboratories*

[3] ISO 15189:2007, *Medical laboratories – Particular requirements for quality and competence*

[4] ISO/IEC 17043:2010, *Conformity assessment — General requirements for proficiency testing*

[5] ISO/IEC Guide 99-12:2007/JCGM 200:2008, *International vocabulary of metrology — Basic and general concepts and associated terms (VIM, 3rd edition)*

[6] EA-4/18, *Guidance on the level and frequency of proficiency testing participation*

[7] EURACHEM - Information leaflet on Pre- and post-analytical proficiency testing, First English edition, 2009-05-14

[8] ILAC Brochure: 2008, *Benefits for laboratories participating in proficiency testing programs*

[9] DIRECTIVE 98/79/EC of the European parliament ad of the council of 27 October 1998 on in vitro diagnostic medical devices, Official Journal of the European Communities.

[10] Eurolab Technical Report 1/2007 – *Measurement uncertainty revisited, Alternative approaches to uncertainty evaluation March 2007*

[11] ISO 21748:2010, *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation*

[12] ISO 13528:2005, *Statistical methods for use in proficiency testing by interlaboratory comparisons*

[13] Analytical Methods Committee - *Robust Statistic Part I & II*. Analyst 114, 1693-1702, (1989)

[14] 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. 11, 373-378, (2006)

[15] Tholen, D.W., *Statistical treatment of proficiency testing data*, Accred. Qual. Assur., 3 (1998), 362-366

[16] The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure Appl. Chem., 78, 1; 145-196, (2006)

[17] VAM Bulletin - Ellison, S.L.R., LGC, , Issue 33 - Autumn 2005, 21-22

[18] BIPM, IEC, IFCC, ISO, IUPAC, OIML: *Guide to the Expression of Uncertainty in Measurement*. ISO, Geneva, Switzerland, First Edition (1993)

[19] Uhlig, S. and Lischer, P - *Statistically-based performance characteristics in laboratory performance studies*, Analyst, 123 , 167-172 (1998)

[20] Rosario P., Martínez JL, Silván JM . - *Comparison of different statistical methods for evaluation of proficiency test data*, Accred. Qual. Assur. 13:493:499 (2008)

[21] Harms, AV. - *A new approach for proficiency test exercise data evaluation*, Accred. Qual. Assur. 14:253:261 (2009).

[22] Rousseeuw P.J. and Leroy A.M. - *Robust regression and outlier detection*. Wiley, New York (1987)

[23] Analytical Methods Committee - *Robust Statistic Part I & II*. Analyst 114, 1693-1702 (1989)

[24] Meijer, P., DE Maat, M.P.M, Kluft, C., Haverkate, F., and Hans C. 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*, Clinical Chemistry 48:7 1011–1015 (2002)

[25] EN ISO 6508-1:2006, *Metallic materials – Rockwell hardness test. Part 1: Test method (scales A, B, C, D, E, F, G, H, K, N, T)*, Beuth Verlag, Berlin, March 2006

[26] ISO/IEC Guide 43-1: 1997, *Selection and use of proficiency testing scheme by laboratory accreditation bodies. (This standard was withdrawn on 29.01.2010 and revised by ISO/IEC 17043: 2010.)*

[27] EN ISO 6508-2:2006, *Metallic material – Rockwell hardness test. Part 2: Verification and calibration of testing machines (scales A, B, C, D, E, F, G, H, K, N, T)*, Beuth Verlag, Berlin, March 2006

[28] UNCERT COP 14:2000, Gabauer, W., *Manual of Codes of Practice for the Determination of Uncertainties in Mechanical Tests on Metallic Materials, The Estimation of Uncertainties in Hardness Measurements*, Project, No. SMT4-CT97-2165.

[29] Weißmüller, C., Frenz, H. - *Modelle zur Ermittlung der Messunsicherheit in der Härtprüfung; Statistische Auswertung eines Ringversuchs mit 90 Teilnehmern*, In: Tagungsband Werkstoffprüfung, Neu-Ulm 25.-26.11.2004, MAT INFO, Frankfurt 2004.

[30] Weißmüller, C., Frenz, H. - *Berechnung der Messunsicherheit für mechanisch- technologische Prüfverfahren, Seminarunterlagen, unveröffentlicht*, Recklinghausen, 2005.