DRAFT INTERNATIONAL STANDARD

ISO/DIS 13528.2

ISO/TC **69**/SC **6**

Secretariat: **JISC**

Voting begins on:
2013-12-17

Voting terminates on:
2014-03-17

Statistical methods for use in proficiency testing by interlaboratory comparisons

Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires
[Revision of first edition (ISO 13528:2005)]

ICS: 03.120.30

Reference number
ISO/DIS 13528.2:2013(E)

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 13528 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 6, *Measurement methods and results*. The first edition was published in 2005. The second edition provides changes to bring the document into harmony with ISO/IEC 17043:2010, which replaced ISO Guide 43-1:1997. The second edition follows a revised structure, to reflect better the process of the design, analysis, and reporting of proficiency testing schemes. It also eliminated some procedures and added or revised some other sections to be consistent with ISO/IEC 17043 and to provide clarity and correct minor errors. New sections were added for qualitative data and additional robust statistical methods.

0 Introduction

0.1 The purposes of proficiency testing

Proficiency testing involves the use of interlaboratory comparisons to determine the performance of participants (which may be laboratories, inspection bodies, or individuals) for specific tests or measurements, and to monitor their continuing performance. There are a number of typical purposes of proficiency testing, as described in the Introduction to ISO/IEC 17043:2010. These include the evaluation of laboratory performance, the identification of problems in laboratories, establishing effectiveness and comparability of test or measurement methods, the provision of additional confidence to laboratory customers, validation of uncertainty claims, and the education of participating laboratories. The statistical design and analytical techniques applied must be appropriate for the stated purpose(s).

0.2 Rationale for scoring in proficiency testing schemes

A variety of scoring strategies is available and in use for proficiency testing. Although the detailed calculations differ, most schemes compare the participant's deviation from an assigned value with a numerical criterion which is used to decide whether or not the deviation represents cause for concern. The strategies used for value assignment and for choosing a criterion for assessment of the participant deviations are therefore critical. In particular, it is important to consider whether the assigned value and criterion for assessing deviations should be independent of participant results, or should be derived from the results submitted. In this Standard, both strategies are provided for. However, attention is drawn to the discussion that will be found in sections 7 and 8 of the advantages and disadvantages of choosing assigned values or criteria for assessing deviations that are not derived from the participant results. It will be seen that in general, choosing assigned values and assessment criteria independently of participant results offers advantages. This is particularly the case for the criterion used to assess deviations from the assigned value – such as the standard deviation for proficiency assessment or an allowance for measurement error – for which a consistent choice based on suitability for a particular end use of the measurement results is especially useful.

0.3 ISO 13528 and ISO/IEC 17043

ISO 13528 provides support for the implementation of ISO/IEC 17043 – particularly, to the requirements for the statistical design, validation of proficiency test items, review of results, and summary statistics. Annex B of ISO/IEC 17043 briefly describes the general statistical methods that are used in proficiency testing schemes. This International Standard is intended to be complementary to the design requirements and Annex B, providing detailed guidance that is lacking in that document on particular statistical methods for proficiency testing.

The definition of proficiency testing in ISO/IEC 17043 is repeated in ISO 13528, with the Notes that describe different types of proficiency testing and the range of complexity of designs that can be used. This Standard cannot specifically cover all purposes, designs, matrices and measurands. The techniques presented in ISO 13528 are intended to be broadly applicable, especially for newly established proficiency testing schemes. It is expected that statistical techniques used for a particular proficiency testing scheme will evolve as the scheme matures; and the scores, evaluation criteria, and graphical techniques will be refined to better serve the specific needs of a target group of participants, accreditation bodies, and regulatory authorities.

ISO 13528 also applies guidance from a harmonized protocol for the proficiency testing of chemical analytical laboratories [29], but is intended for use with all measurement methods and qualitative identifications. This revision of ISO 13528:2005 contains most of the information from the first edition, extended as necessary by the previously referenced documents and the extended scope of ISO/IEC 17043, which includes proficiency testing for individuals and inspection bodies, and Annex B, which includes considerations for qualitative and ordinal results. Some procedures from the first edition have not been carried forward to this edition; due to experience which indicates those techniques are no longer considered to be appropriate.

This Standard presents statistical techniques that are consistent with other International Standards, particularly those of TC69 SC6, notably the ISO 5725 series of standards on *Accuracy: trueness and precision*. The techniques are also intended to reflect techniques from other international standards, where appropriate, and are intended to be consistent with ISO/IEC Guide 98-3 (GUM) and ISO/IEC Guide 99 (VIM).

0.4 Statistical expertise

ISO/IEC 17043:2010 requires that in order to be competent, a proficiency testing provider shall have access to statistical expertise (clause 4.4.1.4) and shall authorize specific personnel to conduct statistical analysis (clause 4.2.4). Neither ISO/IEC 17043 nor this document can specify further what that necessary expertise is. For some applications an advanced degree in statistics is useful, but usually the needs for expertise can be met by individuals with technical expertise in other areas, who are familiar with basic statistical concepts and common techniques. If an individual is charged with statistical design and/or analysis, it is very important that this person has experience with interlaboratory comparisons, even if that person has an advanced degree in statistics. Conventional advanced statistical training often does not include exercises with interlaboratory comparisons, and the unique causes of measurement error that occur in proficiency testing can seem obscure. The guidance in this International Standard cannot provide all the necessary expertise to consider all applications, and cannot replace the experience gained by working with interlaboratory comparisons.

0.5 Computer software

Computer software that is needed for statistical analysis of proficiency testing data can vary greatly, ranging from simple spread sheet arithmetic for small proficiency testing schemes using known reference values to sophisticated statistical software used for statistical methods reliant on iterative calculations or other advanced numerical methods. Most of the techniques in this International Standard can be accomplished by conventional spread sheet applications, perhaps with customised analysis for a particular scheme or analysis; some techniques will require computer applications that are freely available (at the time of publication). In all cases, the user should verify the accuracy of their calculations, especially when special routines have been entered by the user. However, even when the techniques in this International Standard are appropriate and correctly implemented by adequate computer applications, they cannot be applied without attention from an individual with technical and statistical expertise that is sufficient to identify and investigate anomalies that can occur in any round of proficiency testing.

Statistical methods for use in proficiency testing

1 Scope

This International Standard provides detailed descriptions of statistical methods for organizers to use to design proficiency testing schemes and to analyse the data obtained from those schemes, and provides recommendations on the interpretation of proficiency testing data by participants in such schemes and by accreditation bodies.

This International Standard can be applied to demonstrate that the measurement results obtained by laboratories, inspection bodies, and individuals meet specified criteria for acceptable performance.

This International Standard is applicable to proficiency testing where the results reported are either quantitative or qualitative observations on test items.

NOTE The procedures in this Standard may also be applicable to the assessment of expert opinion where the opinions or judgments are reported in a form which may be compared objectively with an independent reference value or a consensus statistic. For example; when classifying test items into known categories by inspection - or in determining by inspection whether test items arise, or do not arise, from the same original source - and the classification results are compared objectively.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies. In the case of differences between normative references on the use of terms, definitions in ISO 3534 parts 1-2 apply.

ISO 3534-1, Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms

ISO 3534-2, Statistics — Vocabulary and symbols — Part 2: Applied statistics

ISO 5725-1, Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions

ISO/IEC 17043, Conformity assessment — General requirements for proficiency testing

ISO/IEC Guide 99, International vocabulary of metrology — Basic and general concepts and associated terms (VIM)

ISO/IEC Guide 98-3, Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-2, ISO 5725-1, ISO/IEC 17043, ISO/IEC Guide 99, ISO Guide 34, and the following apply. Mathematical symbols are listed in Annex A.

3.1
interlaboratory comparison
Organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions

3.2
proficiency testing
Evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons

NOTE For the purposes of this International Standard, the term “proficiency testing” is taken in its widest sense and includes, but is not limited to:

a) quantitative scheme — where the objective is to quantify one or more measurands for each proficiency test item;

b) qualitative scheme — where the objective is to identify or describe one or more qualitative characteristics of the proficiency test item;

c) sequential scheme — where one or more proficiency test items are distributed sequentially for testing or measurement and returned to the proficiency testing provider at intervals;

d) simultaneous scheme — where proficiency test items are distributed for concurrent testing or measurement within a defined time period;

e) single occasion exercise — where proficiency test items are provided on a single occasion;

f) continuous scheme — where proficiency test items are provided at regular intervals;

g) sampling — where samples are taken for subsequent analysis and the purpose of the proficiency testing scheme includes evaluation of the execution of sampling; and

h) data interpretation — where sets of data or other information are furnished and the information is processed to provide an interpretation (or other outcome).

3.3
assigned value
x_{pt}
Value attributed to a particular property of a proficiency test item

3.4
standard deviation for proficiency assessment
SDPA and σ_{pt}
Measure of dispersion used in the evaluation of results of proficiency testing

NOTE 1 This can be interpreted as a target standard deviation for a population of laboratories that is competent to perform a particular measurement procedure.

NOTE 2 The standard deviation for proficiency assessment applies only to ratio and differential scale results.

NOTE 3 Not all proficiency testing schemes evaluate proficiency based on the dispersion of results.

[revised from ISO/IEC 17043]

3.5
measurement error
Measured quantity value minus a reference quantity value (ISO/IEC Guide 99)

3.6
maximum permissible measurement error
Limit of error
δ_E
extreme value of measurement error, with respect to a known reference quantity value, permitted by specifications or regulations for a given measurement, measuring instrument, or measuring system (ISO/IEC Guide 99)

3.7
***z* score**
Standardized measure of performance, calculated using the participant result, assigned value and the standard deviation for proficiency assessment

NOTE: can be modified (*z'* score) by combining the uncertainty of the assigned value with the SDPA

3.8
zeta score
ζ
Standardized measure of performance, calculated using the participant result, assigned value and the combined standard uncertainties for the result and the assigned value.

3.9
normalized error score
E_n
Standardized measure of performance, calculated using the participant result, assigned value and the combined expanded uncertainties for the result and the assigned value

3.10
proportion of allowed limit score
P_A
Standardized measure of performance, calculated using the participant result, assigned value and the criterion for measurement error in a proficiency test

NOTE: for single results, performance can be expressed as the deviation from the assigned value (D or D%).

3.11
action signal
Indication of a need for action arising from a proficiency test result

EXAMPLE: A *z* score in excess of 2 is conventionally taken as indicative of a need to investigate possible causes; a *z* score in excess of 3 may be taken as indicating a need for corrective action

3.12
consensus value
Value derived from a collection of results in an interlaboratory comparison

NOTE the phrase 'consensus value' is typically used to describe estimates of location and dispersion derived from participant results in a proficiency test round, but may also be used to refer to value derived from results of a specified subset of such results or, for example, from a number of expert laboratories.

3.13
outlier
A member of a set of values which is inconsistent with other members of that set. (from ISO 5725-1)

NOTE 1 An outlier can arise by chance from the expected population, originate from a different population, or be the result of an incorrect recording or other gross error.

NOTE 2 Many schemes use the term outlier to designate a result that generates an action signal. This is not the intended use of the term. While outliers will usually generate action signals, it is possible to have action signals from results that are not outliers.

3.14
participant
Laboratory, organization, or individual that receives proficiency test items and submits results for review by the proficiency testing provider.

3.15
proficiency test item
Sample, product, artefact, reference material, piece of equipment, measurement standard, data set or other information used to assess participant performance in proficiency testing.

NOTE In most instances, proficiency test items meet the ISO Guide 34 definition of "reference material" (3.18)

3.16
proficiency testing provider
Organization which takes responsibility for all tasks in the development and operation of a proficiency testing scheme.

3.17
proficiency testing scheme
Proficiency testing designed and operated in one or more rounds for a specified area of testing, measurement, calibration or inspection.

NOTE A proficiency testing scheme might cover a particular type of test, calibration, inspection or a number of tests, calibrations or inspections on proficiency test items.

3.18
reference material
RM
Material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process. (ISO Guide 34:2009)

NOTE 1 RM is a generic term

NOTE 2 Properties can be quantitative or qualitative (e.g., identity of substances or species)

NOTE 3 Uses may include the calibration of a measurement system, assessment of a measurement procedure, assigning values to other materials, and quality control.

3.19
certified reference material
CRM
Reference material characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability. (ISO Guide 34:2009)

NOTE The concept of value includes qualitative attributes such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities.

4 General Principles

4.1 General requirements for statistical methods

4.1.1 The statistical methods used shall be fit for purpose and statistically valid. Furthermore, any statistical assumptions on which the methods or design are based must be stated in the design or in a written description of the scheme, and these assumptions shall be demonstrated to be reasonable.

4.1.2 The proficiency testing provider shall provide participants with a description of the calculation methods used, an explanation of the general interpretation of results and a statement of any limitations relating to interpretation. This shall be available either in each report for each round or in a separate summary of procedures that is available to participants.

4.1.3 The proficiency testing provider shall ensure that all software is adequately validated.

4.2 General model for participant results

4.2.1 For quantitative results, the general model is given in equation (1).

$$x_i = x_{pt} + \varepsilon_i \qquad (1)$$

With x_i = proficiency test result from participant i
x_{pt} = assigned value
$\boldsymbol{\varepsilon}_i$ = measurement error for participant i, distributed according to a relevant model

NOTE 1 a common model is for $\varepsilon_i \sim N(0, \sigma^2)$ with mean 0 and variance known or unknown. However ε_i could be assumed to follow other statistical distributions.

NOTE 2 the basis of performance evaluation with z scores and σ_{pt} is that in an "idealized" population of competent laboratories, the interlaboratory standard deviation would be σ_{pt} or less.

NOTE3 This model differs from the model in ISO 5725, in that it does not include the laboratory bias term B_i. The reason for this change is that the objective of proficiency testing is to evaluate the fitness of the result, as it would typically be reported to a customer. Therefore the size of the deviation from the assigned value is of interest, and there is no need to evaluate laboratory bias.

4.2.2 For ordinal or qualitative results, other models may be appropriate, or there could be no statistical model.

4.3 General approaches for the evaluation of performance

4.3.1 There are three different general approaches for evaluating performance in a proficiency testing scheme. These approaches are used to meet different purposes for the proficiency testing scheme. The approaches are listed below:

a) performance evaluated by comparison with externally derived criteria;
b) performance evaluated by comparison with other participants;
c) performance evaluated by comparison with claimed measurement uncertainty.

4.3.2 The general approaches can be applied differently for the assigned value and the evaluation interval; for example when the assigned value is the robust mean of participant results, and the evaluation interval is a predetermined σ_{pt} or δ_E ; similarly, in some situations the assigned value can be a reference value, but σ_{pt} can be a robust standard deviation of participant results. In approach c) using uncertainty, the assigned value is typically an appropriate reference value.

4.3.3 The statistical design and data analysis techniques shall be consistent with the stated objectives for the scheme.

5 Guidelines for the statistical design of proficiency testing schemes

5.1 Introduction

Proficiency testing is concerned with the assessment of participant performance and as such does not specifically address bias or precision (although these can be assessed with specific designs). The performance of the participants is undertaken through the statistical evaluation of their results following the measurements or interpretations they make on the proficiency test items. Performance is often expressed in the form of scores which allow consistent interpretation across a range of measurands and can allow results for different measurands to be compared on an equal basis. Performance scores are typically derived by comparing the difference between a reported participant result and an assigned value with an allowable deviation or with an estimate of the uncertainty of the difference. Examination of the performance scores over multiple measurands or rounds of a proficiency testing scheme can provide information on whether individual laboratories show evidence of consistent systematic effects ("bias") or poor long term precision.

The following Sections 5-10 give guidance on the design of quantitative proficiency testing schemes and on the statistical treatment of results, including the calculation and interpretation of various performance scores. Considerations for qualitative proficiency testing schemes (including ordinal schemes) are given in Section 11.

5.2 Basis of a statistical design

5.2.1 According to ISO/IEC 17043, clause 4.4.4.1, the statistical design "shall be developed to meet the objectives of the scheme, based on the nature of the data (quantitative or qualitative including ordinal and categorical), statistical assumptions, the nature of errors, and the expected number of results". Therefore proficiency testing schemes with different objectives and with different sources of error could have different designs.

EXAMPLE 1: a proficiency testing scheme to compare a participant's result against a pre-determined reference value and within limits that are specified before the round begins;

EXAMPLE 2: a proficiency testing scheme to compare participant results with combined results from a group in the same round, and limits that are specified before the round begins;

EXAMPLE 3: a proficiency testing scheme to compare participant results with combined results from a group in the same round, and limits determined by the variability of participant results;

EXAMPLE 4: a proficiency testing scheme to compare participants' results with the assigned value, using the participant's own measurement uncertainty.

NOTE Proficiency testing schemes with other (perhaps secondary) objectives could also require specific statistical designs, such as when an objective is to compare performance of different measurement procedures.

5.2.2 There are various types of data used in proficiency testing, including quantitative, nominal (categorical), and ordinal. Among the continuous variables, some results might be on an interval scale (a scale with an arbitrary zero); or a relative, or ratio scale (on a scale with a true zero). For some measurements on a continuous scale, only a discrete and discontinuous set of values can be realized (for example, sequential dilutions); however, in many cases these results can be treated by techniques that are applicable to continuous variables.

5.2.3 Proficiency testing schemes may be used for other purposes in addition to the above, as discussed in section 0.1 and in ISO/IEC 17043. The design must be appropriate for all stated purposes.

5.3 Considerations for the statistical distribution of results

5.3.1 Statistical assumption for distribution

5.3.1.1 ISO/IEC 17043 (4.4.4.2) requires that statistical analysis techniques are consistent with the statistical assumptions for the data. Most common analysis techniques for proficiency testing assume that a set of results from competent participants will be approximately normally distributed, or at least unimodal and reasonably symmetric (after transformation if necessary). A common additional assumption is that the distribution of results from competent measurements is mixed (or 'contaminated') with results from a population of erroneous values which may generate outliers. Usually, the scoring interpretation relies on the assumption of normality, but only for the underlying assumed distribution for competent participants.

5.3.1.2 It is usually not necessary to verify that results are normally distributed, but it is important to verify approximate symmetry, at least visually. If this is not possible then the proficiency testing provider should use techniques that are robust to asymmetry (see Annex C).

5.3.1.3 When the distribution expected for the proficiency testing scheme is not sufficiently close to a symmetric normal distribution (with contamination by outliers), the proficiency testing provider should select data analysis methods that take due account of the asymmetry expected and that are resistant to outliers, and scoring methods that also take due account of the expected distribution for results from competent participants. This may include

— transformation to provide approximate symmetry;

— methods of estimation that are resistant to asymmetry;

— methods of estimation that incorporate appropriate distributional assumptions (for example, maximum likelihood fitting with suitable distribution assumptions and, if necessary, outlier rejection).

EXAMPLE 1: Results based on dilution, such as for quantitative microbiological counts or for immunoassay techniques, are often distributed according to the logarithmic normal distribution, and so a logarithmic transformation is appropriate as the first step in analysis.

EXAMPLE 2: Counts of small numbers of particles, which may be distributed according to a Poisson distribution, and therefore the evaluation interval may be determined using a table of Poisson probabilities, based on the average count for the group of participants.

5.3.1.4 In some areas of calibration, participant results may follow statistical distributions that are described in the measurement procedure; these defined distributions should be considered in any evaluation protocol.

5.3.2 Reasonableness of assumptions

According to ISO/IEC 17043 section 4.4.4.2, the proficiency testing provider must state the reasons for any statistical assumptions and demonstrate that the assumptions are reasonable. This demonstration could be based on the observed data, results from previous rounds of the scheme, or the technical literature.

NOTE The demonstration of the reasonableness of a distribution assumption is less rigorous than the demonstration of the validity of that assumption.

5.4 Considerations for small numbers of participants

5.4.1 Appropriate methods

The statistical design for a proficiency testing scheme shall consider the minimum number of participants that are needed to meet the objectives of the design, and state alternative approaches that will be used if the minimum number is not achieved (ISO/IEC 17043:2010, clause 4.4.4.3 c)). Statistical methods that are appropriate for large numbers of participants may not be appropriate with limited numbers of participants. Concerns are that statistics determined from small numbers of participant results may not be sufficiently reliable, and a participant could be evaluated against an inappropriate comparison group.

NOTE When there are few participants in a proficiency testing scheme, the recommendations of the IUPAC/CITAC Technical Report: *Selection and use of proficiency testing schemes for a limited number of participants* [**22**] should be considered. In brief, the IUPAC/CITAC report recommends that the assigned value should be based on reliable independent measurements; for example by use of a certified reference material, independent assignment by a calibration or national metrology institute, or by gravimetric preparation. The report further states that the standard deviation for proficiency assessment may not be based on the observed dispersion among participant results for a single round.

5.4.2 Factors affecting minimum number of participants

The minimum number of participants needed for the various statistical methods will depend on a variety of situations:

a) The statistical methods used (a particular robust method or outlier removal);
b) The experience of the participants with the particular proficiency testing scheme;
c) The experience of the proficiency testing provider with the matrix, measurand, methods, and group of participants;
d) Whether the intent is to determine the assigned value or the standard deviation (or both).

Further guidance on techniques for small numbers is provided in Annex D.1.

5.5 Guidelines for choosing the reporting format

5.5.1 Reporting results for proficiency testing

It is a requirement of ISO/IEC 17043 (clause 4.6.1.2), that proficiency testing providers instruct participants to carry out measurements and report results on proficiency test items in the same way as for the majority of routinely performed measurements, except in special circumstances.

This requirement can, in some situations, make it difficult to obtain an accurate assessment of participants' precision and trueness, or competence with a measurement procedure. The proficiency testing provider should adopt a consistent reporting format for the proficiency testing scheme but should, where possible, use units familiar to the majority of participants and choose a reporting format that minimises transcription and other errors. This may include automated warning of inappropriate units when participants are known to report routinely in units other than those required by the scheme.

NOTE Transcription errors in collation of results by the proficiency testing provider can be substantially reduced or eliminated by the use of electronic reporting systems that permit participants to enter their own data directly.

5.5.2 Considerations for replicated observations

If a proficiency testing scheme requires replicate measurements on test items, the participant should be required to report all replicates. This can occur, for example, if an objective is to evaluate a participant's precision on known replicate test items, or when a measurement procedure requires separate reporting of multiple observations. In these situations the proficiency testing provider may also need to ask for the participant's mean (or other estimate of location) and uncertainty to assist data analysis by the proficiency testing provider.

5.5.3 Considerations for censored results

5.5.3.1 Where conventional reporting practice is to report results as 'less than' or 'greater than' a limit (such as a calibration interval or a quantitation limit) and where numerical results are required for scoring, the proficiency testing provider needs to consider how the results will be processed. In these cases, the proficiency testing provider should either adopt validated data treatment and scoring procedures that accommodate censored data (see Annexes C and E.1), or require participants to report the numerical value of the result either in place of, or in addition to, the conventional reported value.

NOTE Performance evaluations should be based on results that a participant would routinely report to customers, and not on results that may be outside the interval in which measurements are routinely reported by the participant.

5.5.3.2 When consensus statistics are used, it may not be possible to evaluate performance if the number of censored values is large enough that a robust method is affected by the censoring. In circumstances where the number of censored results is sufficient to affect a robust method, then the results should be evaluated using statistical methods which allow unbiased estimation in the presence of censored data [**19**], or the results should not be evaluated. When in doubt about the effect of the procedure chosen, the proficiency testing provider should calculate summary statistics and performance evaluations with both all of the various statistical procedures, and investigate the importance of any difference(s).

5.5.3.3 Where a censored result is removed or modified for data analysis, the participant's performance should still be evaluated according to the criteria used for all participants in the scheme.

Annex E.1 has an example of some analysis approaches for censored data. This example shows robust consensus statistics with three different approaches: a) with the censored values removed; b) with the values retained but the '<' sign removed; and c) with the results replaced with half of the limit value.

5.5.4 Considerations for number of reported digits

5.5.4.1 Usually, the number of significant digits to report will be determined by the design of the proficiency testing scheme. When specifying numbers of significant digits to be reported, the rounding error should be negligible compared to the expected variation between participants.

5.5.4.2 Where the number of digits reported under routine measurement conditions has an appreciable adverse effect on data treatment by the proficiency testing provider (for example, where measurement procedures require reporting to a small number of significant digits), the provider may specify the number of digits to be reported.

NOTE For example, a measurement procedure might specify reporting to 0,1 g, leading to a large proportion (>50 %) of identical results and in turn compromising the calculation of robust means and standard deviations. The provider may then require participants to report to two or three decimal places to obtain sufficiently reliable estimates of location and scale.

5.5.4.3 In some situations, correct reporting is part of the determination of competence of the participant, and the number of significant digits and decimal places can vary. If it is allowed that different participants will report results using different numbers of significant digits, the proficiency testing provider needs to take this into consideration when generating any consensus statistics (such as the assigned value and standard deviation for proficiency assessment).

6 Guidelines for the initial review of proficiency testing items and results

6.1 Homogeneity and stability of proficiency test items

6.1.1 The proficiency testing provider shall ensure that batches of proficiency test items are sufficiently homogeneous and stable for the purposes of the proficiency testing scheme. Criteria shall ensure that inhomogeneity and instability of proficiency test items do not adversely affect the evaluation of performance. The assessment of homogeneity and stability should use one of the following approaches:

a) Experimental studies as described in Annex B or alternative experimental methods that provide equivalent assurance of homogeneity and stability;

b) Experience with the behaviour of closely similar materials in previous rounds of the proficiency testing scheme, verified as necessary for the current round.

c) *A posteriori* assessment of participant data for evidence of consistency with previous rounds, or evidence of change with reporting time or unexpected dispersion attributable to inhomogeneity or instability.

These situations can be justified on a case-by-case basis, using appropriate statistical techniques and technical justification.

NOTE For example, if previous rounds of a proficiency testing scheme used proficiency test items that were tested and demonstrated to be sufficiently homogeneous and stable, and with the same participants as in previous rounds, then if an interlaboratory standard deviation in the current round is not greater than the standard deviation in previous rounds, that is evidence of sufficient homogeneity and stability in the current round.

6.1.2 For calibration proficiency testing schemes where the same artefact is used by multiple participants, the proficiency testing provider shall assure stability throughout the round, or have procedures to identify and account for instability through the progression of a round. This should include consideration of tendencies for particular artefacts and measurands, such as drift. Where appropriate, the assurance of stability should consider the effects of multiple shipments of the same artefact.

6.1.3 Ideally, all measurands (or properties) should be checked for homogeneity and stability, but it is possible to test a subset of properties. If a property is not checked, the proficiency testing provider should have a documented explanation for the correlation between any property that is not checked and a property that is checked. The measurands that are checked should be sensitive to sources of inhomogeneity or instability in the processing of the proficiency test item. Some examples are:

a) when the measurement is a proportion, a characteristic that is a small proportion can be more difficult to homogenize and so be more sensitive in a homogeneity check;

b) if a proficiency test item is heated during processing, then choose a measurand that is sensitive to uneven heating;

c) if a measured property can be affected by settling, precipitation, or other time-dependent effects during the preparation of proficiency test items, then this property should be checked across filling order.

6.2 Considerations for different measurement methods

6.2.1 When all participants are expected to report a value for the same measurand, the assigned value for proficiency testing should normally be the same for all participants. However, when participants are allowed to choose their own measurement method, it is possible that a single assigned value for each analyte or property will not be appropriate for all participants. This can occur, for example, when different measurement methods provide results that are not comparable. In this case, the proficiency testing provider may assign a different assigned value for each measurement method.

Examples:

(a) medical testing where different approved measurement methods are known to respond differently to the same test material and use different reference ranges for diagnosis

(b) operationally defined methods, such as leachable toxic metals in soils, for which different standard methods are available and are not expected to be directly compared, but where the proficiency testing scheme specifies the measurand without reference to a specific test method.

6.2.2 The need for different assigned values for subsets of participants should be considered in the design of the proficiency testing scheme (for example, to make provision for reporting of specific methods) and should also be considered when reviewing data for each round

6.3 Blunder removal

6.3.1 ISO/IEC 17043:2010 section B.2.5 and the IUPAC Harmonized Protocol recommend removing obvious blunders from a data set at an early stage in an analysis, prior to use of any robust procedure or any test to identify statistical outliers. Generally, these results would be treated separately (such as contacting the participant). It can be possible to correct some blunders, but this should only be done according to an approved policy and procedure.

Occasionally (especially in proficiency testing that is offered for the first time) a single participant (or multiple participants) will make several such blunders in reporting, perhaps due to misunderstanding of instructions or report forms. In such cases all results from that participant should be removed from the analysis, including any results that may not be obvious blunders.

NOTE Obvious blunders, such as reporting results in incorrect units or switching results from different proficiency test items, occurs in most rounds of proficiency testing, and these results only impair the performance of subsequent statistical methods.

6.3.2 If there is any doubt about whether a result is a blunder, it should be retained in the data set and subjected to subsequent treatment, as described in sections 6.4 to 6.6.

6.4 Visual review of data

6.4.1 The first step in any data analysis is visual review of the data, conducted by a person who has adequate technical and statistical expertise. This check is to confirm the expected distribution of results, and to identify anomalies, or unanticipated sources of variability. For example, a bimodal distribution might be evidence of a mixed population of results caused by different methods, contaminated samples or poorly worded instructions. In this situation, the concern should be resolved before proceeding with analysis or evaluation.

NOTE 1 A histogram is a useful and widely available review procedure, to look for a distribution that is unimodal and symmetric, and to identify unusual outliers. However the intervals used for combining results in a histogram are sensitive to numbers of results and cut points, and so can be difficult to create. A kernel density plot is often more useful for identifying possible bimodalities or lack of symmetry (see section 10.3).

NOTE 2 Other review techniques can be useful, such as a cumulative distribution plot or a stem-and-leaf diagram can be useful. Graphical methods for data review are illustrated in Annex E.1.

6.4.2 When it is not feasible to conduct visual review of all data sets of interest, there shall be a procedure to warn of unexpected variability in a dataset; for example by reviewing the uncertainty of the assigned value compared to the evaluation criteria, or by comparison with previous rounds.

Annexes E.2 and E.3 have comprehensive data sets and analyses to cover many techniques in this document. Visual review of data is part of these examples.

6.5 Robust statistical methods

6.5.1 Robust statistical methods provide procedures that can be thought of as describing a central part of the distribution of results from competent participants, but without requiring the identification of specific values as outliers and excluding them from subsequent analyses. Typical statistics used are based on the median and the central 50% of results - these are measures of the center and spread of the data, similar to the mean and standard deviation (respectively). In general, robust methods are preferred to methods that delete results labelled as outliers.

6.5.2 The median, scaled median absolute deviation (MAD), and normalized IQR are allowed as simple estimators, but can lead to having a large number of unacceptable results. The "Q Method" is useful especially for situations where a large proportion (>20%) of results can be discrepant, or data cannot be reliably reviewed by experts. Algorithm A transforms the original data by a process called winsorisation for estimators of mean and standard deviation based on the median/MAD method.

Details for robust procedures are described in Annex C.

6.5.3 The choice of statistical methods is the responsibility of the proficiency testing provider. The robust mean and standard deviation can be used for various purposes, of which the evaluation of performance is just one. Robust means and standard deviations may also be used as summary statistics for different groups of participants or for specific methods.

NOTE Robust methods can give misleading results if they are applied to data sets that are markedly skewed or multimodal.

Annexes E.2 and E.3 have comprehensive examples illustrating the use of a variety of robust statistical techniques presented in Annex C.

6.6 Outlier techniques for individual results

6.6.1 Outlier tests may be used either to support visual review for anomalies or to provide a degree of resistance to extreme values when calculating summary statistics. If outlier detection techniques are used, they should be clearly defined and demonstrated to be in compliance with ISO 16269-4 [**10**], or other appropriate techniques. This requires valid assumptions about the nature of the distribution of the population of results, and controls for the probability of incorrect identification (that is, the likelihood that a result actually belongs as a member of the population). Ideally, the chosen method will have defined breakdown points and efficiencies under various possible conditions. It should also be recognized that the validity of any parametric outlier removal is based on the validity of an assumption for the underlying distribution. If an outlier procedure is based on an underlying normal distribution for results from competent participants, that assumption should be demonstrated to be reasonable.

6.6.2 Where outlier rejection is part of a data handling procedure, and a result is removed as an outlier, the participant's performance should still be evaluated according to the criteria used for all participants in the scheme.

NOTE 1 Outliers among reported values are often identified by employing Grubbs' test for outliers, as given in ISO 5725-2. Evaluation in this procedure is applied using the standard deviation of all participants including potential outliers. Therefore this procedure should be applied when the performance of participants is consistent with expectations from previous rounds and there are a small number of outliers (one or two outliers on each side of the mean). Conventional tables for Grubbs' procedure assume a single application for a possible outlier (or 2) in a defined location, not unlimited sequential application. If the Grubbs' tables are applied sequentially, the Type I error probabilities for the tests may not apply.

NOTE 2 When replicate results are returned or identical proficiency test items are included in a round, it is also common to use Cochran's test for repeatability outliers, also described in ISO 5725-2.

NOTE 3 Outliers may also be identified by robust or nonparametric techniques; for example if a robust mean and standard deviation are calculated, values deviating from the robust mean by more than 3 times the robust standard deviation might be identified as outliers.

7 Determination of the assigned value and its standard uncertainty

7.1 Choice of method of determining the assigned value

7.1.1 Five ways of determining the assigned value x_{pt} are described in sections 7.3 to 7.7. The choice between these methods is the responsibility of the proficiency testing provider.

NOTE Sections 7.3-7.6 are closely similar to approaches used to determine the property values of certified reference materials described in ISO Guide 35.

7.1.2 Alternative methods for determining the assigned value and its uncertainty may be used provided that they have a sound statistical basis and that the method used is described in the documented plan for the proficiency testing scheme, and fully described to participants. Regardless of the method used to determine the assigned value, it is always appropriate to check the validity of the assigned value for that round of proficiency testing. This is discussed in section 7.8.

7.1.3 Approaches for determining qualitative assigned values are discussed in section 11.3.

7.2 Determining the uncertainty of the assigned value

7.2.1 The *Guide to the expression of uncertainty in measurement* (ISO/IEC Guide 98) gives guidance on the evaluation of measurement uncertainties. ISO Guide 35 provides guidance on the uncertainty of the assigned value for certified property values, which can be applied for many proficiency testing scheme designs.

7.2.2 A general model for the assigned value and its uncertainty is described in equations (2) and (3):

The model for the assigned value can be expressed as follows;

$$x_{pt} = x_{char} + \delta x_{hom} + \delta x_{trans} + \delta x_{stab} \qquad (2)$$

where

x_{pt} denotes the assigned value;
x_{char} denotes the property value obtained from the characterization (determination of assigned value);
δx_{hom} denotes an error term due to the difference between proficiency test items;
δx_{trans} denotes an error term due to instability under transport conditions;
δx_{stab} denotes an error term due to instability during the period of proficiency testing.

The associated model for the uncertainty of the assigned value can be expressed as follows:

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2} \qquad \text{with:} \qquad (3)$$

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

u_{char} = standard uncertainty due to characterization;

u_{hom} = standard uncertainty due to differences between proficiency test items;

u_{trans} = standard uncertainty due to instability caused by transport of proficiency test items;

u_{stab} = standard uncertainty due to instability during the period of proficiency testing.

NOTE 1 Covariance between sources of uncertainty, or negligible sources, may lead to a different model for specific applications. Any of the components of uncertainty can be zero or negligible, in some situations.

NOTE 2 When σ_{pt} is calculated as the standard deviation of participant results, the uncertainty components due to inhomogeneity, transport, and instability are reflected in the variability of participant results. In this case the uncertainty of the assigned value, as described in sections 7.3-7.7, are sufficient.

NOTE 3 The proficiency testing provider is normally expected to ensure that changes related to instability or incurred in transport are negligible compared to the standard deviation for proficiency assessment; that is, to ensure that δx_{trans} and δx_{stab} are zero. Where this requirement is met, u_{stab} and u_{trans} may be set to zero.

7.2.3 There can be bias in the assigned value that is not accounted for in the above expression. This shall, where possible, be considered in the design for the proficiency testing scheme. If there is an adjustment for bias in the assigned value, the uncertainty of this adjustment must be included in the evaluation of the uncertainty of the assigned value.

7.3 Formulation

7.3.1 General

7.3.1.1 The proficiency test item can be prepared by mixing materials with different known levels of a property in specified proportions, or by adding a specified proportion of a substance to a base material. In this case, the assigned value x_{pt} is derived by calculation from the masses of properties used. This approach is especially valuable when individual proficiency test items are prepared in this way, and it is the proportion of the properties that is to be determined.

7.3.1.2 Reasonable care should be taken to ensure that:

a) the base material is effectively free from the added constituent, or that the proportion of the added constituent in the base material is accurately known;
b) the constituents are mixed together homogeneously (where this is required);
c) all significant sources of error are identified (e.g., it is not always realized that glass absorbs mercury compounds, so that the concentration of an aqueous solution of a mercury compound can be altered by its container);
d) there is no adverse interaction between the constituents and the matrix;
e) the behaviour of proficiency test items containing added material may differ from the proficiency test items routinely tested. For example, pure materials added to a natural matrix often extract more readily than the same substance occurring naturally in the material. If there is a concern about this happening, the proficiency testing provider should assure the suitability of the proficiency test items for the methods that will be used.

7.3.1.3 When formulation gives samples in which the addition is more loosely bonded than in typical materials, or in a different form, it may be preferable to use another approach to prepare proficiency test items.

7.3.1.4 Determination of the assigned value by formulation is a one case of a general approach for characterization of certified reference materials described by ISO Guide 35, where a single laboratory determines an assigned value using a primary measurement method. Other uses of a primary method by a single laboratory can be used to determine the assigned value for proficiency testing.

7.3.2 Standard uncertainty of the assigned value from formulation

When the assigned value is calculated from the formulation of the test material, the standard uncertainty for the characterization (u_{char}) is estimated by combination of uncertainties using an appropriate model. For example, in chemical analyses the uncertainties will usually be those associated with gravimetric and volumetric measurements and the purity of any spiking materials. The standard uncertainty of the assigned value ($u(x_{pt})$) is then calculated according to equation (3).

7.4 Certified reference material

7.4.1 General

When a proficiency test item is a certified reference material (CRM), its certified property value x_{CRM} is used as the assigned value x_{pt}.

Limitations of this approach are:

- it can be expensive to provide every participant with a unit of a certified reference material;

- CRMs are often processed quite heavily to ensure long-term stability, which may compromise the commutability of the samples.
- a CRM may be known to the participants making it important to conceal the identity of the proficiency test item.

7.4.2 Standard uncertainty of the assigned value from a certified reference material

When a certified reference material is used as the proficiency test item, the standard uncertainty of the assigned value is derived from the information on the uncertainty of the property value provided on the certificate. The certificate information should include the components in equation (3), and have an intended use appropriate for the purpose of the proficiency testing scheme.

7.5 Results from one laboratory

7.5.1 General

In this approach, the proficiency test items are prepared and made ready for distribution to the participants. An appropriate number of the samples are then selected at random and tested using a certified reference material as a calibrator, in one competent laboratory, using a suitable measurement method. The assigned value x_{pt} of the test material is then derived from a calibration against the reference values of the CRM. This approach applies to many types of proficiency testing, including the approach for many calibration artefacts. The approach assumes that the CRM is commutable for all measurement methods used by participants.

When a CRM is not available, reference values can be obtained from any appropriate source; the reference value should have defined metrological traceability and uncertainty that is appropriate for the proficiency testing scheme.

7.5.2 Procedure to determine an assigned value by comparison with a CRM

A CRM described is used to determine the assigned value for a proficiency test item which is another, similar, reference material. This determination requires a series of tests to be carried out, in one laboratory, on samples of the two aggregates, using the same measurement method, and under repeatability conditions. Let

x_{CRM} is the assigned value for the CRM
x_{pt} is the assigned value for the proficiency test item
d_i is the difference between the average results for the proficiency test item and the CRM on the i^{th} samples
$\bar{d}$ is the average of the differences d_i

Then,

$$x_{pt} = x_{CRM} + \bar{d} \quad (4)$$

NOTE Where commutability is a concern, care should be taken to ensure that the CRM or RM used is sufficiently commutable with the proficiency test items for the purposes of the scheme.

7.5.3 Standard uncertainty of the assigned value from a single laboratory

When the assigned value is derived from the results of a series of tests on that proficiency test item and a CRM, the standard uncertainty of characterization is derived from the uncertainty of the measurement used for value assignment. This approach allows the assigned value to be established in a manner that is metrologically traceable to the certified value of the CRM, with a standard uncertainty that can be calculated from equation (3).

The standard uncertainty of the assigned value of the RM may be calculated as:

$$u(x_{pt}) = \sqrt{u_{CRM}^2 + u_d^2} \quad (5)$$

The example in Annex E.4 illustrates how the required uncertainty may be calculated in the simple case when the assigned value of a proficiency test item is established by direct comparison with a single CRM.

7.6 Consensus value from expert laboratories

7.6.1 General

This is the same approach as used in ISO Guide 35 for use of interlaboratory comparisons to characterize a CRM. Proficiency test items are prepared first and made ready for distribution to the participants. Some of these samples are then selected at random and analysed by a group of experts using a protocol that specifies the numbers of samples and replicates and any other relevant conditions. An expert laboratory is required to provide an estimate of the uncertainty for any reported result.

7.6.2 Standard uncertainty of the assigned value from the consensus value from expert laboratories

The variance (squared standard uncertainty) associated with the assigned value is determined as the average of the variances associated with the results from expert laboratories, together with the variance components due to homogeneity, transport, and stability as described in equation (3). When each of p experts reports a measurement x_i on the test material together with an estimate $u(x_i)$ of the standard uncertainty of the measurement, and the assigned value x_{pt} is calculated as a robust average using Algorithm A, or other method, the standard uncertainty for characterization of the assigned value is estimated as:

$$u_{char} = \frac{1{,}25}{p} \times \sqrt{\sum_{i=1}^{p} u(x_i)^2} \qquad (6)$$

The standard uncertainty of the assigned value ($u(x_{pt})$) is then calculated according to equation (3).

The limitations of this approach are that there may be an unknown bias in the results of the group of experts, and the claimed uncertainties may not be reliable. It is likely that p will be small, so this approach should be used with caution.

If all uncertainties were estimated correctly, then all results from all expert laboratories should agree within their stated uncertainties. Disagreement within the stated uncertainties should be taken as a warning sign and an additional uncertainty contribution should be added to account for this difference. This additional uncertainty can be estimated as $SD(x_i)/\sqrt{p}$, where $SD(x_i)$ is the standard deviation of the means from i=1…p expert laboratories.

An example of using combined results from expert laboratories is provided in Annex E.5.

7.7 Consensus value from participant results

7.7.1 General

7.7.1.1 With this approach, the assigned value x_{pt} for the proficiency test item used in a round of a proficiency testing scheme is the location estimate (e.g., robust mean, median, or arithmetic mean) of the results reported by participants in the round, calculated using an appropriate procedure in accordance with the design, as described in Annex C. Techniques described in sections 6.2-6.6 should be used to confirm that sufficient agreement exists, before combining results.

7.7.1.2 In some situations, the proficiency testing provider may wish to use a subset of participants determined to be reliable, by some pre-defined criteria, such as accreditation status or on the basis of prior performance. The techniques of this section apply to those situations, including considerations for group size.

7.7.1.3 Other calculation methods may be used in place of those in Annex C, provided that they have a sound statistical basis and the report describes the method that is used.

7.7.1.4 The consensus value approach may be particularly useful with an operationally defined measurement method, provided that the method is standardized.

7.7.1.5 The limitations of this approach are that:

a) there may be insufficient agreement among the participants;

b) the consensus value may include unknown bias due to the general use of faulty methodology and this bias will not be reflected in the standard uncertainty of the assigned value;

c) the consensus value could be biased due to the effect of bias in methods that are used to determine the assigned value;

d) It may be difficult to determine the metrological traceability of the consensus value. While the result is always traceable to the results of the individual laboratories, a clear statement of traceability beyond that can only be made when the proficiency testing provider has complete information about the calibration standards used by all of the participants contributing to the consensus value.

7.7.2 Standard uncertainty of the assigned value from the consensus value of participant results

7.7.2.1 The standard uncertainty of the assigned value will depend on the procedure used. If a fully general approach is needed, the proficiency testing provider should consider the use of resampling techniques ("bootstrapping") to estimate a standard error for the assigned value. References [**16** and **17**] give details of bootstrapping techniques.

7.7.2.2 When the assigned value is derived as a robust average calculated using procedures in Annex C.1 or C.3, the standard uncertainty of the assigned value x_{pt} may be estimated as:

$$u(x_{pt}) = 1{,}25 \times \frac{s^*}{\sqrt{p}} \quad (7)$$

where

s^* is the robust standard deviation of the results calculated using procedures in Annex C.1 or C.3. (Here a "result" for a participant is the average of all their measurements on the proficiency test item.)

NOTE 1 In this model, where the assigned value and robust standard deviation are determined from participant results, the uncertainty of the assigned value can be assumed to include the effects of uncertainty due to inhomogeneity, transport, and instability.

NOTE 2 The factor 1,25 is based on the standard deviation of the median, or the efficiency of the median as an estimate of the mean, in a large set of results drawn from a normal distribution. It is appreciated that the efficiency of more sophisticated robust methods can be much greater than that of the median, justifying a correction factor smaller than 1,25. However, this factor has been recommended because proficiency testing results typically are not strictly normally distributed, and contain unknown proportions of results from different distributions ('contaminated results'). The factor of 1,25 is considered to be a conservative (high) estimate, to account for possible contamination. Proficiency testing providers may be able to justify using a smaller factor, or a different equation, depending on experience and the robust procedure used.

An example of using a value from participants is provided in Annexes E.2 and E.3.

7.8 Comparison of the assigned value with an independent reference value

7.8.1 When the methods described in 7.7 are used to establish the assigned value (x_{pt}), and where a reliable independent estimate (denoted x_{ref}) is available, for example from knowledge of preparation or from a reference value, the consensus value x_{pt} should be compared with x_{ref}.

When the methods described in 7.3 to 7.6 are used to establish the assigned value, the robust average x^* derived from the results of the round should be compared with the assigned value after each round of a proficiency testing scheme. The difference is calculated as x_{diff} = (x_{ref}-x_{pt}) (or (x^*-x_{pt})) and the standard uncertainty of the difference ($u(x_{ref}$-$x_{pt})$) is estimated as:

$$u_{diff} = \sqrt{u(x_{ref})^2 + u(x_{pt})^2} \quad (8)$$

where

$u(x_{ref})$ is the uncertainty of the reference value for comparison; and
$u(x_{pt})$ is the uncertainty of the assigned value.

An example of a comparison of a reference value with a consensus value is included in Annex E.2.

7.8.2 If the difference is more than twice its standard uncertainty, the reason should be investigated. Possible reasons are:

- bias in the measurement method;
- a common bias in the results of the participants;
- failure to appreciate the limitations of the measurement method when using the formulation method described in 7.3;
- bias in the results of the "experts" when using the approaches in sections 7.5 or 7.6; and
- the comparison value and assigned value are not traceable to the same metrological reference.

7.8.3 Depending on the reason for the disagreement, the PT provider should decide whether to evaluate results or not, and (for continuous schemes), whether to amend the design for subsequent schemes. Where the difference is sufficiently large to affect performance assessment or to suggest important bias in the measurement methods used by participants, the difference should be noted in the report for the round. In such cases, the difference should be considered in the design of future studies.

8 Determination of criteria for evaluation of performance

8.1 Approaches for determining evaluation criteria

8.1.1 The basic approach for all purposes is to compare a result on a proficiency test item (x_i) with an assigned value (x_{pt}). For evaluation, the difference is compared to an allowance for measurement error. This comparison is commonly made through a standardized performance statistic (e.g., *z*, *z'*, *zeta*, E_n), as discussed in sections 9.3-9.6. This can also be done by comparing the difference with a defined criterion (D or D% compared to δ_E) as discussed in 9.2. An alternative approach to evaluation is to compare the difference with a participant's claim for uncertainty of their result combined with the uncertainty of the assigned value (E_n and *zeta*).

8.1.2 If a regulatory requirement or a fitness for purpose goal is given as a standard deviation it may be used directly as σ_{pt}. If the requirement or goal is for a maximum permissible measurement error, that criterion can be divided by the action limit to obtain σ_{pt}. A prescribed limit of error could be used directly as δ_E for use with D or $D\%$.

NOTE For example, if a regulatory criterion is specified as a limit of error and 3,0 is an action limit for evaluation with a *z* score, then the specified criterion is divided by 3,0 to determine σ_{pt}.

8.1.3 When the criterion for evaluation of performance is based on consensus statistics from the current round or previous rounds, then a robust estimate of the standard deviation of participant results is the preferred statistic. When this approach is used it is usually most convenient to use a score such as the *z* score and to set the standard deviation for proficiency assessment (σ_{pt}) to the calculated estimate of the standard deviation.

8.2 By perception of experts

8.2.1 The limit of error or the standard deviation for proficiency assessment may be set at a value that corresponds to the level of performance that a regulatory authority, accreditation body, or the technical experts of the proficiency testing provider believe is reasonable for participants.

8.2.2 A specified limit of error can be transformed into a standard deviation for proficiency assessment by dividing the limit by the number of multiples of the σ_{pt} that are used to define an action signal (or unacceptable result). Similarly, a specified σ_{pt} can be transformed into δ_E.

An example of deriving a value by expert (or regulatory) mandate is provided in Annexes E.2 and E.3.

8.3 By experience from previous rounds of a proficiency testing scheme

8.3.1 The standard deviation for proficiency assessment (σ_{pt}), and the limit of error (δ_E), can be determined by experience with previous rounds of proficiency testing for the same measurand with comparable property values, and where participants use compatible measurement procedures. This is a useful approach when there is no agreement among experts about fitness for purpose. The advantages of this approach are as follows:

— evaluations will be based on reasonable performance expectations;
— the evaluation criteria will not vary from round to round because of random variation or changes in the participant population;
— the evaluation criteria will not vary between different proficiency testing providers, when there are two or more approved proficiency testing providers approved for an area of testing or calibration.

8.3.2 The review of previous rounds should include consideration of performance that is achievable by competent participants, and not affected by new participants or random variation due to, for example, smaller group sizes or other factors unique to a particular round. Determinations can be made subjectively by examination of previous rounds for consistency, or objectively with averages or with a regression model that adjusts for the value of the measurand. The regression equation might be a straight line, or could be curved [**30**]. Standard deviations and relative standard deviations should be considered, with selection based on which is more consistent across the appropriate range of measurand levels. Expectation for maximum permissible measurement error can also be obtained in this manner.

An example of deriving a value from experience of previous rounds is provided in Annex E.6.

8.4 By use of a general model

8.4.1 General considerations

8.4.1.1 The value of the standard deviation for proficiency assessment can be derived from a general model for the reproducibility of the measurement method. This method has the advantage of objectivity and consistency across measurands, as well as being empirically based. Depending on the model used, this approach could be considered a special case of a fitness for purpose criterion.

8.4.1.2 Any expected standard deviation chosen by a general model must be reasonable. If very large or very small proportions of participants are assigned action or warning signals, the proficiency testing provider should ensure that this is consistent with the purpose of the proficiency testing scheme.

8.4.2 Example: Horwitz curve

One common general model for chemical applications was described by Horwitz [16] and modified by Thompson [**28**]. This approach gives a general model for the reproducibility of analytical methods that may be used to derive the following expression for the reproducibility standard deviation:

$$\sigma_R = \begin{cases} 0{,}22c & for \quad c < 1{,}2 \times 10^{-7} \\ 0{,}02c^{0{,}8495} & for \quad 1{,}2 \times 10^{-7} \le c \le 0{,}138 \\ 0{,}01c^{0{,}5} & for \quad c > 0{,}138 \end{cases} \qquad (9)$$

where

c is the concentration of the chemical species to be determined in mass fraction $0 \le c \le 1$.

NOTE This model is empirical, based on observations from collaborative trials. The σ_R values are the expected upper limits of interlaboratory variability when the collaborative trial had no significant problems.

An example of deriving a value from the modified Horwitz model is provided in Annex E.7.

8.5 Using the repeatability and reproducibility standard deviations from a previous collaborative study of precision of a measurement method

8.5.1 When the measurement method to be used in the proficiency testing scheme is standardized, and information on the repeatability (σ_r) and reproducibility (σ_R) of the method is available, the standard deviation for proficiency assessment (σ_{pt}) may be calculated using this information, as follows:

$$\sigma_{\text{pt}} = \sqrt{\sigma_R^2 - \sigma_r^2(1 - 1/m)} \tag{10}$$

where

m is the number of replicate measurements each participant is to perform in a round of the scheme.

8.5.2 When the repeatability and reproducibility standard deviations are dependent on the average value of the test results, functional relations should be derived by the methods described in ISO 5725-2. These relations should then be used to calculate values of the repeatability and reproducibility standard deviations appropriate for the assigned value that is to be used in the proficiency testing scheme.

8.5.3 For the techniques above to be valid, the collaborative study must have been conducted according to the requirements of ISO 5725-2 or equivalent procedure.

An example is presented in Annex E.8.

8.6 From data obtained in the same round of a proficiency testing scheme

8.6.1 With this approach, the standard deviation for proficiency assessment, σ_{pt}, is used to evaluate the performance of participants in a round of a proficiency testing scheme; it is derived from the combined measurement results reported by the participants in the same round. This is usually the robust standard deviation of the results reported by all the participants, calculated using a technique listed in Annex C. In general, evaluation with D or $D\%$ and using δ_{E} are not appropriate in these situations, however P_A can still be used as a standardized score, for comparison across measurands.

8.6.2 The main advantages of this approach are simplicity and conventional acceptance due to successful use in many situations. This approach may be the only feasible approach, especially when there is no national or international reference.

8.6.3 There are several disadvantages with this approach:

a) The value of σ_{pt} may vary substantially from round to round, making it difficult for a participant to use values of the *z* score to look for trends that persist over several rounds.

b) Standard deviations can be unreliable when the number of participants in the proficiency testing scheme is small or when results from different methods are combined. For example, if p=20, the standard deviation for normally distributed data can vary by about +-30% from its true value from one round to the next.

c) Using dispersion measures derived from the data will lead to an approximately constant proportion of apparently acceptable scores. Generally poor performance will not be detected by inspection of the scores, and generally good performance will result in good participants receiving poor scores.

d) There is no useful interpretation in terms of suitability for any end use of the results.

Examples of using participant data are provided in the comprehensive example in Annexes E.2.

8.7 Monitoring interlaboratory agreement

8.7.1 As a check on the performance of the participants, and to assess the benefit of the proficiency testing scheme to the participants, it is recommended that the proficiency testing provider should apply a procedure to monitor interlaboratory agreement, to track changes in performance and ensure the reasonableness of statistical procedures.

8.7.2 The results obtained in each round of a proficiency testing scheme should be used to calculate estimates of the reproducibility standard deviations of the measurement method (and repeatability, if available), using the robust methods described in Annex C. These estimates should be plotted on graphs sequentially or as a time-series, together with values of the repeatability and reproducibility standard deviations obtained in precision experiments from ISO 5725-2 (if available), and/or σ_{pt}, if techniques in sections 8.2 to 8.4 are used.

8.7.3 These graphs should then be examined by the proficiency testing provider. If the graphs show that the precision values obtained in a specific proficiency testing scheme differ by a factor of two or more from the values obtained in the precision experiment, then the proficiency testing provider should investigate why agreement in this specific test was worse than before. Similarly, a trend towards better or worse precision values should trigger an investigation for the most likely causes.

9 Calculation of performance statistics

9.1 General considerations for determining performance

9.1.1 Statistics used for determining performance shall be consistent with the objective(s) for the proficiency testing scheme.

NOTE Performance statistics are most useful if the statistics and their derivation are understood by participants and other interested parties.

9.1.2 There are advantages to having standardized scores that are easily reviewed across measurand levels and different rounds of a proficiency testing scheme.

9.1.3 Performance statistics are meaningful only when the participant results have been reviewed and are determined to be consistent with the design of the proficiency testing scheme. That is, when there is no evidence of deterioration of the proficiency test item, or of a mixture of populations of participants, or of severe violations of any statistical assumptions about the nature of the data.

9.1.4 In general, it is not appropriate to use evaluation methods that intentionally classify a fixed proportion of results as generating an ‘action signal’.

9.2 Limiting the uncertainty of the assigned value

9.2.1 If the standard uncertainty $u(x_{pt})$ of the assigned value is large in comparison with the performance evaluation criterion, then there is a risk that some participants will receive action and warning signals because of inaccuracy in the determination of the assigned value, not because of any cause of the participant. For this reason, the standard uncertainty of the assigned value shall be determined and shall be reported to participants (see ISO/IEC 17043:2010, 4.4.5 and 4.8.2).

If the following criterion is met, then the uncertainty of the assigned value is negligible and need not be included in the interpretation of the results of the proficiency test.

$$u(x_{pt}) < 0{,}3\sigma_{pt} \quad \text{or} \quad u(x_{pt}) < 0{,}1\delta_E \text{ (when } \delta_E = 3\sigma_{pt}) \qquad (11)$$

9.2.2 If this criterion is not met, then the proficiency testing provider should consider the following, ensuring any action taken remains consistent with the agreed performance assessment policy for the proficiency testing scheme.

a) Select a method for determining the assigned value such that its uncertainty meets the criterion in equation (11).

b) Use the uncertainty of the assigned value in the interpretation of the results of the proficiency testing scheme (see sections 9.5 on the *z'* score, or 9.6 on zeta scores, or 9.7 on E_n scores).

c) If the assigned value is derived from participant results, and the large uncertainty arises from differences between identifiable sub-populations of participants, report separate values and uncertainties for each sub-population (for example, participants using different measurement methods).

NOTE The IUPAC Harmonized Protocol [**29**] describes a specific procedure for detecting bimodality, based on an inspection of a kernel density plot with a specified bandwidth.

d) Inform the participants that the uncertainty of the assigned value is not negligible, and evaluations could be affected.

If none of a)-d) apply, then the participants shall be informed that no reliable assigned value can be determined and that no performance scores can be provided.

The techniques presented in this section are demonstrated in Annex E.3.

9.3 Estimates of deviation (measurement error)

9.3.1 General

Let x_i represent the result (or the average of the replicates) reported by a participant i for the measurement of a property of the proficiency test item in one round of a proficiency testing scheme. Then a simple measure of performance of the participant can be calculated as the difference between the result x_i and the assigned value x_{pt}:

$$D_i = x_i - x_{pt} \tag{12}$$

D_i can be interpreted as the measurement error for that result, to the extent to which the assigned value can be considered a conventional or reference quantity value.

The difference D_i may be expressed in the same units as the assigned value or as a percentage difference, calculated as:

$$D_i\% = 100(x_i - x_{pt})/x_{pt} \tag{13}$$

9.3.2 Interpretation of differences

9.3.2.1 The difference D_i or $D_i\%$ is usually compared with a criterion δ_E based on fitness for purpose or with experience from previous rounds; the criterion is noted here as δ_E, an allowance for measurement error. If $-\delta_E < D_i < \delta_E$ then the performance is considered to be 'acceptable' (or 'no signal'), i.e., there is not sufficient evidence of measurement error that exceeds the criterion. (The same criterion applies for $D_i\%$, depending on the expression of δ_E.)

9.3.2.2 δ_E is closely related to σ_{pt} as used for *z* scores (see 9.4), when σ_{pt} is determined by fitness for purpose or expectations from previous rounds. The relation is determined by the evaluation criterion for *z* scores. For example, if $z \geq 3$ creates an action signal then $\delta_E = 3\,\sigma_{pt}$, or equivalently $\sigma_{pt} = \delta_E/3$. Various expressions of δ_E are conventional in proficiency testing for medical applications and in performance specifications for measurement methods and products.

9.3.2.3 The advantage of D as a performance statistic and δ_E as a performance criterion is that participants have an intuitive understanding of these statistics since they are tied directly to measurement error and are common as criteria to determine fitness for purpose. Disadvantages are that it is not conventional for proficiency testing in many countries or fields of measurement; and that D is not standardized, to allow simple scanning of reports for action signals in proficiency testing schemes with multiple analytes or where fitness for purpose criteria can vary by level of the measurand.

NOTE Use of D and $D\%$ generally assumes symmetry of the distribution of participant results in the sense that the acceptable range is $-\delta_E < D < \delta_E$.

9.3.3 Standardized Difference score

For purposes of comparison across measurand levels, where fitness for purpose criteria can vary; or for combination across rounds or across measurands, D and $D\%$ can be transformed into standardized scores that show the differences relative to the performance criteria for the measurands. To do this, calculate the "Percentage of Allowed Deviation" (P_A) for every result as follows:

$$P_{Ai} = (D_i / \delta_E) \times 100\% \quad (14)$$

Therefore $P_A \geq 100\%$ or $P_A \leq -100\%$ indicates an action signal (or 'unacceptable performance').

P_A scores can be compared across levels and different rounds, or tracked in charts. These scores are similar in use and interpretation to *z* scores that have a common evaluation criterion such as $z \leq -3$ or $z \geq 3$ for action signals.

Variations of this statistic are commonly used, particularly in medical applications, where there is usually a higher frequency of proficiency testing and a large number of analytes.

It may be appropriate to use the absolute value of P_A to reflect consistently acceptable (or unacceptable) results relative to the assigned value.

9.4 *z* scores

9.4.1 General

The *z* score for a proficiency test result x_i is calculated as:

$$z_i = (x_i - x_{pt})/\sigma_{pt} \quad (15)$$

Where

x_{pt} is the assigned value, and

σ_{pt} is the standard deviation for proficiency assessment.

9.4.2 Interpretation of *z* scores

9.4.2.1 The conventional interpretation of *z* scores is as follows (see ISO/IEC 17043 section B.4.1.1 c)):

- A result that gives $|z| \leq 2{,}0$ is considered to be acceptable;
- A result that gives $2{,}0 < |z| < 3{,}0$ is considered to give a warning signal;
- A result that gives $|z| \geq 3{,}0$ is considered to be unacceptable (or action signal).

Participants should be advised to check their measurement procedures following warning signals in case they indicate an emerging or recurrent problem.

NOTE 1 In some applications, proficiency testing providers use 2,0 as an action signal for *z* scores.

NOTE 2 The choice of criterion σ_{pt} should normally be made so as to permit the above interpretation, which is widely used for proficiency assessment and is also closely similar to familiar control chart limits.

NOTE 3 The justification for the use of the limits of 2,0 and 3,0 for *z* scores is as follows. Measurements that are carried out correctly are assumed to generate results that can be described (after transformation if necessary) by a normal

distribution with mean x_{pt} and standard deviation σ_{pt}. *z* scores will then be normally distributed with a mean of zero and a standard deviation of 1,0. Under these circumstances only about 0,3 % of scores would be expected to fall outside the range -3,0 ≤ z ≤ 3,0 and only about 5 % would be expected to fall outside the range -2,0 ≤ z ≤ 2,0. Because the probability of *z* falling outside ±3,0 is so low, it is unlikely that action signals will occur by chance when no real problem exists, so it is likely that there is an identifiable cause for an anomaly when an action signal is given.

NOTE 4 The assumption on which this interpretation is based applies only to a hypothesized distribution of competent laboratories and not on any assumption about the distribution of the observed results. No assumption needs to be made about the observed results themselves.

NOTE 5 If the actual true interlaboratory variability is smaller than σ_{pt} then the probabilities of misclassification are reduced.

NOTE 6 When the standard deviation for proficiency assessment is fixed by either of the methods described in 8.2 or 8.4, it may differ substantially from the (robust) standard deviation of results, and the proportions of results falling outside ±2,0 and ±3,0 may differ considerably from 5% and 0,3% respectively.

9.4.2.2 The proficiency testing provider shall determine the appropriate rounding interval for *z* scores, based on the number of significant digits for the result, and for the assigned value and the standard deviation for proficiency testing. The rounding interval shall be included in the design information available to participants.

9.4.2.3 When the standard deviation of participant results is used as σ_{pt} and proficiency testing schemes involve very large numbers of participants, the proficiency testing provider may wish to check the normality of the distribution, using actual results or *z* scores. At the other extreme, when there are only a small number of participants, there may be no action signal given. In this case, graphical methods that combine scores over several rounds may provide more useful indications of the performance of the participants than the results of individual rounds.

9.5 z' scores

9.5.1 General

When there is concern about the uncertainty of an assigned value $u(x_{pt})$, for example when $u(x_{pt}) > 0{,}3\sigma_{pt}$, then the uncertainty can be taken into account by expanding the denominator of the performance score. This statistic is called a z' score and is calculated as follows (with notation as in section 9.4):

$$z_i' = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u(x_{pt})^2}} \quad (16)$$

NOTE 1 When the assigned value x_{pt} is a consensus value from participant results, the estimate of the mean is correlated with individual participant scores. Individual results can have an impact on both a robust mean and standard deviation - the correlation for an individual participant depends on the weighting given to that participant in the combined statistic. For this reason, scores including the uncertainty of the assigned value without including an allowance for correlation represent under-estimates of the score that would result if the covariance were included. The under-estimation is not serious if the assigned value uncertainty is small, and when robust methods are used it is least serious for the outermost participants most likely to receive adverse scores. Therefore equation (16) can usually be used without adjustment for correlation.

NOTE 2 D_i and $D_i\%$ scores can also be modified to consider the uncertainty of the assigned value with the following formula to expand δ_E to δ_E'

$$\delta_E' = \sqrt{\delta_E^2 + U(x_{pt})^2} \quad (17)$$

Where

$U(x_{pt})$ is the expanded uncertainty of the assigned value x_{pt} calculated with coverage factor k=2.

9.5.2 Interpretation of *z′* scores

z′ scores can be interpreted in the same way as *z* scores (see 9.4) and using the same critical values of 2,0 and 3,0, depending on the design for the proficiency testing scheme. Similarly, D_i and $D_i\%$ scores would then be compared with δ_E' (see 9.3).

9.5.3 Use of *z′* scores

Comparison of the formulae for the *z* score and the z′ score in 9.4 and 9.5 shows that the *z′* scores for a round of a proficiency testing scheme will always be smaller than the corresponding *z* scores by a constant factor of

$$\frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2+u(x_{pt})^2}}$$

When the guideline for limiting the uncertainty of the assigned value in 9.2.1 is met, this factor will fall in the range:

$$0{,}96 < \frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2+u(x_{pt})^2}} < 1{,}00$$

Thus, in this case, the *z′* scores will be nearly identical to the *z* scores, and it may be concluded that the uncertainty of the assigned value is negligible for the evaluation of performance.

When the guideline in 9.2.1 for the uncertainty of the assigned value is not met, the difference in magnitude of the *z′* scores and *z* scores may be such that some *z* scores exceed the critical values of 2,0 or 3,0 and so give "warning signals" or "action signals", whereas the corresponding *z′* scores do not exceed these critical values and so do not give signals.

In general, for situations when the assigned value and/or σ_{pt} is not determined from participant results, *z'* may be preferred because when the criterion in 9.2.1 is met the difference between *z* and *z'* will be negligible.

9.6 Zeta scores (ζ)

9.6.1 General

9.6.1.1 Zeta scores can be useful when an objective for the proficiency testing scheme is to evaluate a participant's ability to have results be close to the assigned value within their claimed uncertainty.

With notation as in 9.4, the ζ scores are calculated as:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u(x_i)^2+u(x_{pt})^2}} \qquad (18)$$

Where

$u(x_i)$ is the participant's own estimate of the standard uncertainty of its result x_i, and
$u(x_{pt})$ is the standard uncertainty of the assigned value x_{pt}.

9.6.1.2 When combined participant results are used as the assigned value (x_{pt}), then x_{pt} is correlated with individual participant results. The correlation for an individual participant depends on the weighting given to that participant in the assigned value, and to a lesser extent, in the uncertainty of the assigned value. For this reason, scores including the uncertainty of the assigned value without including an allowance for correlation represent under-estimates of the score that would result if the covariance were included. The under-estimation is not serious if the uncertainty of the assigned value is small; when robust methods are used it is least serious for the outermost participants most likely to receive adverse scores. Equation (18) may therefore be used without adjustment for correlation.

NOTE 1 ζ scores differ from E_n scores by using standard uncertainties $u(x_i)$ and $u(x_{pt})$, rather than expanded uncertainties $U(x_i)$ and $U(x_{pt})$. ζ scores above 2 or below -2 may be caused by systematically biased methods or by a poor

estimation of the measurement uncertainty by the participant. ζ scores therefore provide a rigorous assessment of the complete result submitted by the participant.

NOTE 2 When a ζ score is used alone, it can be interpreted only as a test of whether the participant's uncertainty is consistent with the particular observed deviation and cannot be interpreted as an indication of the fitness for purpose of a particular laboratory's results. Determination of fitness for purpose could be done separately (for example, by the participant or by an accrediting body) by examining the deviation $(x\text{-}x_{pt})$ or the combined standard uncertainties in comparison with a target uncertainty.

9.6.2 Interpretation of ζ scores

9.6.2.1 ζ scores may be used when there is an effective system in operation for validating participants' own estimates of the standard uncertainties of their results. Such scores can be interpreted using the same critical values of 2,0 and 3,0 as for *z* scores, or perhaps with multiples from the participant's coverage factor used when estimating expanded uncertainty. However, such a system is the responsibility of the participants and any other party that is interested in the participant's reported uncertainty. It is not usually the responsibility of the proficiency testing provider to assess the validity of uncertainties reported by participants, but a useful guideline for assessment is suggested in section 9.8.

9.6.2.2 ζ scores can be used in conjunction with *z* scores, as an aid for improving the performance of participants, as follows. If a participant obtains *z* scores that repeatedly exceed the critical value of 3,0, they may find it of value to examine their test procedure step by step and derive an uncertainty evaluation for that procedure. The uncertainty evaluation will identify the steps in the procedure where the largest uncertainties arise, so that the participant can see where to expend effort to achieve an improvement. If the participant's ζ scores also repeatedly exceed the critical value of 3,0, it implies that the participant's uncertainty evaluation does not include all significant sources of uncertainty (i.e., they are missing something important). Conversely, if a participant repeatedly obtains *z* scores ≥ 3 but ζ scores < 2, this demonstrates that the participant may have assessed the uncertainty of their results accurately but that their results do not meet the performance expected for the proficiency testing scheme. This may be the case, for example, for a participant who uses a screening method in measurement procedures where the other participants apply quantitative methods. No action is needed if the participant deems that the uncertainty of its results is sufficient.

9.7 E_n scores

9.7.1 General

E_n scores can be useful when an objective for the proficiency testing scheme is to evaluate a participant's ability to have results be close to the assigned value within their claimed uncertainty. This statistic is conventional for proficiency testing in calibration, but it can be used for other types of proficiency testing.

This performance statistic is calculated as:

$$E_n = \frac{x_i - x_{pt}}{\sqrt{U(x_i)^2 + U(x_{pt})^2}} \tag{19}$$

where

x_{pt} is the assigned value determined in a reference laboratory;

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

$U(x_i)$ is the expanded uncertainty of a participant's result x_i.

NOTE Direct combination of expanded uncertainties is not consistent with the requirement of ISO/IEC Guide 98 and is not equivalent to the calculation of a combined expanded uncertainty unless both the coverage factors and the effective degrees of freedom are identical for $U(x_i)$ and $U(x_{pt})$

9.7.2 Interpretation of E_n

E_n scores shall be interpreted with caution (see NOTEs below), because it is a ratio of two separate (but related) performance measures. The numerator is the deviation of the result from the assigned value, and

has an interpretation discussed in section 9.3. The denominator is a combined expanded uncertainty that should not be larger than the deviation in the numerator, if the participant has determined $U(x_i)$ correctly and if the proficiency testing provider has determined $U(x_{pt})$ correctly. Therefore, scores of $E_n \geq 1{,}0$ or $E_n \leq -1{,}0$ could indicate a need to review the uncertainty estimates, or to correct a measurement issue; similarly $-1{,}0 < E_n < 1{,}0$ should be taken as an indicator of successful performance only if the uncertainties are valid and the deviation (x-x_{pt}) is smaller than needed by the participant's customers.

NOTE 1 E_n scores have no probabilistic interpretation unless both the coverage factors and the effective degrees of freedom are identical for $U(x_i)$ and $U(x_{pt})$ and the confidence associated with the coverage factor is known

NOTE 2 While the interpretation of E_n scores can be difficult, that should not prevent their use. Incorporating information on uncertainty into the interpretation of results of proficiency testing results can play a major role in improving the participants' understanding of measurement uncertainty and its evaluation.

9.8 Interpretation of participant uncertainties in testing

9.8.1 General

With increasing application of ISO/IEC 17025 there is better understanding of measurement uncertainty. The use of laboratory evaluations of uncertainty in performance evaluation has been common in proficiency testing schemes in different areas of calibration, such as with the E_n scores, but it has not been common in proficiency testing for testing laboratories. The zeta scores described in section 9.6, and E_n scores in section 9.7, are options for evaluation of results against the claimed uncertainty.

Examples of the analysis of data when uncertainties are reported in Annexes E.2 and E.3.

9.8.2 Guidelines for interpreting uncertainties reported by participants

9.8.2.1 Some proficiency testing providers have recognized the usefulness of asking laboratories to report the uncertainty of results in proficiency testing. This can be useful even when the uncertainties are not used in scoring. There are several purposes for gathering such information:

a) Accreditation bodies can assure that participants are reporting uncertainties that are consistent with their scope of accreditation;

b) Participants can review their reported uncertainty along with those of other participants, to assess consistency (or not) and thereby gain an opportunity to identify whether the uncertainty is not counting all relevant components, or is over-counting some components.

c) Proficiency testing can be used to confirm claims of uncertainty (Introduction to ISO/IEC 17043), and this is easiest when the uncertainty is reported with the result.

9.8.2.2 Where $u(x_{pt})$ meets the criterion in 9.2.1 then it is unlikely that a participant result will have smaller standard uncertainty than this, so $u(x_{pt})$ could be used as a lower limit for screening, called u_{min}. If the assigned value is determined by other methods, the proficiency testing provider may determine other practical screening limits for u_{min}.

NOTE If $u(x_{pt})$ includes variability due to inhomogeneity or instability, the participant's $u(x_i)$ could be smaller than u_{min}.

9.8.2.3 It is also unlikely that any participant's reported standard uncertainty is larger than 1,5 times the robust standard deviation of participants ($1.5s^*$), so this could be used as a practical upper limit for screening reported uncertainties, called u_{max}.

9.8.2.4 If u_{min} or u_{max}, or other criteria, are used to identify aberrant uncertainties, the proficiency testing provider should explain this to participants, and make it clear that a reported uncertainty ($u(x_i)$) can be valid even if it is lower than u_{min} or larger than u_{max}; and when this occurs participants and any interested parties should check the result or the uncertainty estimate. Similarly, a reported uncertainty can be larger than u_{min} and smaller than u_{max}, and still not be valid. These are informative indicators only.

9.8.2.5 Proficiency testing providers may also draw attention to unusually high or low uncertainties based on, for example:

- specified quantiles for the reported uncertainties (for example below the 5^{th} percentile and above the 95^{th} percentile of the reported standard or expanded uncertainties);

- other nonparametric limits, for example the whisker limits conventionally used for box plots (usually set at 1.5 times the interquartile range outside the quartiles);

- limits based on an assumed distribution with scale based on the dispersion of reported uncertainties or on a required measurement uncertainty.

NOTE Since uncertainties are unlikely to be normally distributed, transformation is likely to be necessary when using limits that rely on approximate or underlying normality; for example box plot whisker limits based on the interquartile range have a probabilistic interpretation when the distribution is approximately normal.

9.9 Combined performance scores

9.9.1 General guidelines for combining scores.

9.9.1.1 It is common, within a single round of a proficiency testing scheme, for results to be obtained for more than one proficiency test item or for more than one measurand. In this situation, the results for each proficiency test item and for each measurand should be interpreted as described in 9.3 to 9.7; i.e., the results for each proficiency test item and each measurand should be evaluated separately.

9.9.1.2 There are applications when two or more proficiency test items with specially designed levels are included in a proficiency testing scheme to measure other aspects of performance, such as to investigate repeatability, systematic error, or linearity. For example, two similar proficiency test items may be used in a proficiency testing scheme with the intention of treating them with a Youden plot, as described in 10.5. In such instances, the proficiency testing provider should provide participants with complete descriptions of the statistical design and procedures that are used.

9.9.1.3 The graphical methods described in Section 10 should be used when results are obtained for more than one proficiency test item or for several measurands, provided they are closely related and/or obtained by the same method. These procedures combine scores in ways that do not conceal high values of individual scores, and they may reveal additional information on the performance of participants - such as correlation between results for different measurands - that is not apparent in tables of the individual scores.

9.9.2 Combining performance evaluations.

In proficiency testing schemes that involve a large number of measurands, a count or proportion of the numbers of action and warning signals can be used to allow the participants that obtain one or more such signals to be identified. Participants should also be provided with a report containing detailed results using the methods described in 9.3 to 9.7.

9.9.3 Combined scores

9.9.3.1 For some purposes, proficiency testing providers may wish to combine performance scores across proficiency test items or for related measurands in a proficiency testing scheme. This has proven to be a useful approach for evaluating performance when there are a large number of measurands and proficiency test items. The combined scores could be, for example, a sum or average of squared *z* scores, or an average of P_A scores. Another approach is to award points for accuracy, for example based on *z* score or difference compared to a goal. Similarly some schemes may award points as a penalty for larger error. Commonly, the points accumulated by a participant are combined – perhaps as a sum, or an average, or as a proportion of possible points.

9.9.3.2 Combined scores or award or penalty scores should be used only with caution, because it can be difficult to describe the statistical assumptions underlying the scores. While combined scores for results on different proficiency test items on the same measurand can have expected distributions and can be useful for detecting persistent bias, averaged or summed scores across different measurands on the same or different proficiency test items can conceal bias in results for single measurands. The method of calculation, the interpretation, and the limitations of any combined or penalty scores used shall therefore be made clear to participants.

9.9.3.3 If participant results, penalty points, or award points are combined abd then transformed into z scores for performance evaluation, the distribution of scores should be checked for approximate normality prior to performance evaluation, using the techniques in section 6.4.

10 Graphical methods for describing performance scores from one round of a proficiency test

10.1 Application

The proficiency testing provider should normally use the performance scores obtained in each round of a proficiency testing scheme to prepare graphs such as those described in 10.2 and 10.3. The use of performance scores, such as P_A, *z*, *z'*, *zeta*, or E_n scores in these graphs has the advantage that they can be drawn using standardized axes, thereby simplifying their presentation and interpretation. Graphs should be made available to the participants, enabling each participant to see where their own results fall in relation to those obtained by the other participants. Letter codes or number codes can be used to represent the participants so that each participant is able to identify their own results but not able to determine which participant obtained any other result. The graphs may also be used by the proficiency testing provider and any appropriate accrediting body, to enable them to judge the overall effectiveness of the scheme and to see if there is a need for reviewing the criteria used to evaluate performance.

10.2 Histograms of results or performance scores

10.2.1 General

The histogram is a common statistical tool, and is useful at two different points in the analysis of proficiency testing results. The graph is useful in the preliminary analysis stage, to check whether the statistical assumptions are reasonable, or if there is an anomaly - such as a bimodal distribution, a large proportion of outliers, or unusual skewness that was not anticipated.

Histograms can also be useful in reports for the proficiency testing scheme, to describe the performance scores, or to compare results on, for example, different methods or different proficiency test items. Histograms are particularly useful in individual reports for small or moderate-sized proficiency testing schemes (fewer than 100 participants) to allow participants to assess how their performance compares with other participants, for example, by highlighting a block within a vertical bar to represent a participant's result or, in small proficiency testing schemes (fewer than 50 participants), using individualized plot characters for each participant.

10.2.2 Preparing histograms

Histograms can be prepared using actual participant results or performance scores. Participant results have the advantage of being directly related to the submitted data and can be assessed without further calculation or transformation from the performance score to the measurement error. Histograms based on performance scores have the advantage of relating directly to performance evaluations, and can easily be compared across measurands and rounds of a proficiency testing scheme.

The range and bin size used for a histogram must be determined for each set of data, based on the variability and the number of results. It is often possible to do this based on experience with proficiency testing, but in most situations the groupings will need to be adjusted after the first view. If performance scores are used in the histogram, it is useful to have a scale based on the standard deviation for proficiency assessment and cut points for warning and action signals.

10.2.3 Interpretation

The scale and plot intervals should be chosen so that bimodality can be detected (if it is present), without creating false warnings due to the resolution of measurement results or small numbers of results.

NOTE The appearance of histograms is sensitive to the bin width chosen and to the location of bin boundaries (for constant bin width this is largely dependent on the starting point). If the bin width is too small, the plot will show many small modes; too large and appreciable modes near the main body may not be sufficiently distinct. The appearance of narrow modes and the relative heights of adjacent bars my change appreciably on changing starting position or bin width, especially where the data set is small and/or shows some clustering.

An example of a histogram plot is provided in Annexes E.2 and E.3.

10.3 Kernel density plots

10.3.1 General

A kernel density plot, often abbreviated to 'density plot', provides a smooth curve describing the general shape of the distribution of a data set. The idea underlying the kernel estimate is that each data point is replaced by a specified distribution (typically normal), centred on the point and with a standard deviation σ_k; σ_k is usually called the 'bandwidth'. These distributions are added together and the resulting distribution, scaled to have a unit area, gives a 'density estimate' which can be plotted as a smooth curve.

10.3.2 Construction of a kernel density plot

The following steps may be followed to prepare a kernel density plot. It is assumed that a data set X consisting of p values $x_1, x_2, ..., x_p$ are to be included in the plot. These are usually participant results but may be scores derived from the results.

i) Choose an appropriate bandwidth σ_k. Two options are particularly useful:

a) For general inspection, set $\sigma_k = 0{,}9\, s^*/p^{0.2}$ where s^* is a robust standard deviation of the values $x_1, ..., x_p$ calculated using procedures in Annex C.2 or C.3.

b) To examine the data set for gross modes that are important compared to the criterion for performance assessment, set $\sigma_k = 0.75\sigma_{pt}$ if using z or ζ scores, or $\sigma_k = 0.25\delta_E$ if using D or $D\%$.

Note 1 Option a) above follows Silverman [**25**], which recommends s^* based on the normalised interquartile range (nIQR). Other bandwidth selection rules that provide similar results include that of Scott [**26**], which replaces the multiplier of 0,9 with 1,06. Reference [**27**] describes a near-optimal, but much more complex, method of bandwidth selection. In practice, the differences for visual inspection are slight and the choice depends on software availability.

Note 2 Option b) above follows IUPAC guidance [**29**].

ii) Set a plotting range q_{min} to q_{max} so that $q_{min} \leq \min(x_1, ... x_p) - 3\sigma_k$ and $q_{max} \geq \max(x_1, ... x_p) + 3\sigma_k$.

iii) Choose a number of points n_k for the plotted curve. $n_k = 200$ is usually sufficient unless there are extreme outliers within the range of the plot.

iv) Calculate plotting locations q_1 to q_{n_k} from

$$q_i = q_{min} + (i-1)\frac{(q_{n_k} - q_1)}{n_k - 1} \qquad (20)$$

v) Calculate n_k densities h_1 to h_{n_k} from

$$h_i = \frac{1}{p}\sum_{j=1,p} \phi\left(\frac{x_j - q_i}{\sigma_k}\right) \text{ for } i = 1 \text{ to } i = n_k \qquad (21)$$

where $\phi(.)$ denotes the standard normal density.

vi) Plot h_i against q_i.

Examples of kernel density plots are given in Annexes E.2 and E.3.

Note 1 It may be useful to add the locations of the individual data points to the plot. This is most commonly done by plotting the locations below the plotted density curve as short vertical markers (sometimes called a 'rug'), but may also be done by plotting the data points at the appropriate points along the calculated density curve.

Note 2 Density plots are best done by software. The above stepwise calculation can be done in a spread-sheet for modest data set sizes. Proprietary and freely available statistical software often includes density plots based on similar default bandwidth choices. Advanced software implementations of density plots may use this algorithm or faster calculations based on convolution methods.

10.3.3 Interpretation

The shape of the curve is taken as an indication of the distribution from which the data were drawn. Distinct modes appear as separate peaks. Outlying values appear as separate peaks well separated from the main body of the data.

Note 1 A density plot is sensitive to the bandwidth σ_k chosen. If the bandwidth is too small, the plot will show many small modes; too large and appreciable modes near the main body may not be sufficiently distinct.

Note 2 Like histograms, density plots are best used with moderate to large data sets because small data sets (ten or fewer) may by chance include mild outliers or apparent modes, particularly when a robust standard deviation is used as the basis for the bandwidth.

10.4 Bar-plots of standardized scores

10.4.1 Bar-plots are a suitable method of presenting the performance scores for a number of similar characteristics in one graph. They will reveal if there is any common feature in the scores for a participant, for example if a participant achieves several high *z* scores indicating generally poor performance, that participant may have positive bias.

10.4.2 To prepare a bar-plot, collect the standardized scores into a bar-plot as shown in Figure E.9, in which scores for each participant are grouped together. Other standardized scores, such as $D\%$ or P_A can be plotted for the same purpose.

10.4.3 When replicate determinations are made in a round of a proficiency testing scheme, the results may be used to calculate a graph of precision measures; for example, k statistics as described in ISO 5725-2, or a related measure scaled against the robust average standard deviation such as that defined in Algorithm S (Annex C.4).

10.5 Youden Plot

10.5.1 General

When two similar proficiency test items have been tested in a round of a proficiency testing scheme, the Youden Plot provides a very informative graphical method of studying the results. It can be useful for demonstrating correlation (or independence) of results on different proficiency test items, and for guiding investigations into reasons for action signals.

10.5.2 Preparing a Youden plot.

The graph is constructed by plotting the participant results, or the *z* scores, obtained on one of the proficiency test items against the participant results or *z* scores obtained on the other proficiency test item. Vertical and horizontal lines are typically drawn to create four quadrants of values, to assist interpretation. The lines are drawn at the assigned values or at the medians for the two distributions of results, or drawn at 0 if *z* scores are plotted.

For appropriate interpretation of Youden plots it is important that the two proficiency test items have similar (or identical) levels of the measurand; this is so that the nature of any systematic measurement error is the same in that area of the measuring interval. Youden plots can be useful for widely different levels of a measurand in the presence of consistent systematic error, but they can be deceptive if a calibration error is not consistently positive or negative across the range of measurand levels.

A confidence ellipse, calculated as described in Annex E.9.2, can be useful as an aid to interpretation of the plot. Alternatively, warning and action signal cut points can be plotted as squares; for example squares could be drawn at z=+/-3 on both axes if scores are plotted, or at +/-σ_{pt} or +/-δ_E if measurement results are plotted. The confidence ellipse should be used with caution because it can lead to conflicting interpretation, since both proficiency testing results could be evaluated as “no signal” or “warning”, yet the plotted point could lie outside

the ellipse; or on the other hand it is possible to have “action signals” on both results, yet the point could lie within an ellipse. The interpretation of the ellipses must be explained well to avoid confusion.

10.5.3 Interpretation.

When a Youden Plot is constructed, interpretation is as follows:

a) Inspect the plot for points that are well-separated from the rest of the data. If a participant is not following the test method correctly, so that its results are subject to systematic error, a point will be given far out in the lower left or upper right quadrants. Points far from the others in the upper left and lower right quadrants represent participants whose repeatability is larger than most other participants, whose measurement methods show different sensitivity to the proficiency test item composition type or, sometimes, participants who have accidentally interchanged proficiency test items.

b) Inspect the plot to see if there is evidence of a general relationship between the results for the two proficiency test items (for example, if they lie approximately along a sloped line). If there is evidence of a relationship, then it shows that there is evidence of laboratory bias that affects different samples in a similar way. If there is no apparent visual relationship between results (e.g., points would lie within a circle) than the measurement errors are random in nature. This can be checked with a rank correlation statistic, if the visual examination is not conclusive.

c) Inspect the plot for close groups of participants, either along the diagonals or elsewhere. Clear groups are likely to indicate differences between different methods.

NOTE In studies where all participants use the same measurement method, or plots of results from a single measurement method, if results lie along a line, this is evidence that the measurement method has not been adequately specified. Investigation of the test method may then allow the reproducibility of the method to be generally improved.

An example of a confidence ellipse and a rank correlation test is provided in Annex E.9.

10.6 Plots of repeatability standard deviations

10.6.1 General

When replicate measurements are made by the participants in a round of a proficiency testing scheme, the results may be used to produce a plot to identify any participants whose average and standard deviation are unusual.

10.6.2 Constructing the plot

The graph is constructed by plotting the within-participant standard deviation s_i for each participant against the corresponding average x_i for the participant. Alternatively the range of replicate results can be used instead of the standard deviation. Let

$\bar{X}$ = x^* the robust average of x_1, x_2, ..., x_p, as calculated by Algorithm A

$\bar{S}$ = w^* the robust pooled average of s_1, s_2, ..., s_p, as calculated by Algorithm S

and assume that the data are normally distributed. Under the null hypothesis that there is no difference between participants in the population values of either the participant means or the within-participant standard deviations, the statistic

$$(\sqrt{m}\frac{x_i-\bar{X}}{\bar{s}})^2 + (\sqrt{2(m-1)}\ln\left(\frac{s_i}{\bar{s}}\right))^2 \quad (22)$$

has approximately the χ^2 distribution with 2 degrees of freedom. Hence a critical region with a significance level of 1 % may be drawn on the graph by plotting

$$s = \bar{S}^{\pm\frac{1}{\sqrt{2(m-1)}}\sqrt{\chi^2_{2;0,99}-(\sqrt{m}\frac{x-\bar{X}}{\bar{S}})^2}} \quad (23)$$

on the standard deviation axis against x on the average axis for

$$x = \bar{X} - \bar{S}\sqrt{\frac{\chi^2_{2;0,99}}{m}} \quad to \quad \bar{X} + \bar{S}\sqrt{\frac{\chi^2_{2;0,99}}{m}} \tag{24}$$

NOTE This procedure is based on the Circle Technique introduced by van Nuland [22]. The method described used a simple Normal approximation for the distribution of the standard deviation that could give a critical region containing negative standard deviations. The method given here uses an approximation for the distribution of the standard deviation that avoids this problem, but the critical region is no longer a circle as in the original. Further, robust values are used for the central point in place of simple averages as in the original method.

10.6.3 Interpretation.

The plot can indicate participants with bias that is unusually large, given their repeatability. If there are a large number of replicates, this technique can also identify participants with exceptionally small repeatability. However, because there are usually a small number of replicates, interpretations are rare.

An example of a plot of repeatability standard deviations is provided in Annex E.9.3.

10.7 Split samples

10.7.1 General

Split samples are used when it is necessary to carry out a detailed comparison of two participants, or when proficiency testing is not available and some external verification is needed. Samples of several materials are obtained, representing a wide range of the property of interest, each sample is split into two parts, and each laboratory obtains some number (at least two) of replicate determinations on part of each sample.

On occasion, more than two participants may be involved, in which case one should be treated as a reference, and the others should be compared with it using the techniques described here.

NOTE This type of study is common, but often named differently, such as “paired sample” or “bilateral comparisons”.

10.7.2 Preparation

The data from a split-sample experiment can be used to produce graphs that display the variation between replicate measurements for the two participants and the differences between their average results for each sample. Bivariate plots using the full range of concentrations can have a scale that makes it difficult to identify visually important differences between participants, so plots of the differences or percentage differences between results from the two participants can be more useful. Further analysis will be dependent on deductions made from these graphs.

An example of treating data from split-level design is provided in Annex E.9.4

10.8 Graphical methods for combining performance scores over several rounds of a proficiency testing scheme

10.8.1 Applications

When standardized scores are to be combined over several rounds, the proficiency testing provider may consider preparing graphs, as described in 10.8.2 or 10.8.3. The use of these graphs, in which the scores for several proficiency testing rounds are combined, can allow trends, and other features of the results, to be identified that are not apparent when scores for each round are examined separately.

NOTE The use of “running scores” or “cumulative scores”, in which the scores obtained by a participant are combined over several rounds but not displayed graphically, is not recommended. The participant may have a fault that shows up with the proficiency test item used in one round but not in the others. A running score could hide this fault.

10.8.2 Shewhart control chart for standardized score

10.8.2.1 The Shewhart control chart is an effective method of identifying problems that cause large erratic values of z scores. See ISO 8258 [6] for advice on plotting Shewhart charts and rules for action limits.

10.8.2.2 To prepare this chart, standardized scores, such as z scores or P_A scores, for a participant are plotted as individual points, with action and warning limits set consistent with the design for the proficiency testing scheme; see Annex E.10. When several characteristics are measured in each round, the scores for different characteristics may be plotted on the same graph, but the points for the different characteristics should be plotted using different plotting symbols and/or different colours. When several proficiency test items are included in the same round the scores can be plotted together with multiple points at each time period. Lines joining the mean scores at each time point may also be added to the plot.

10.8.2.3 Conventional rules for interpreting the Shewhart control chart are that an out-of-control signal is given when

a) a single point falls outside the action limits ($\pm$ 3,0 for z scores, or 100% for P_A);

b) two out of three successive points outside either warning limit ($\pm$ 2,0);

c) six consecutive results either positive or negative.

10.8.2.4 When a Shewhart control chart gives an out-of-control signal, the participant should investigate possible causes.

NOTE The standard deviation for proficiency assessment σ_{pt} is not usually the standard deviation of the differences ($x_i - x_{\text{pt}}$), so the probability levels that are usually associated with the action and warning limits of a Shewhart control chart may not apply.

An example of a Shewhart control chart for z-scores is provided in Annex E.10

10.8.3 Plots of standardized scores against assigned values

When the level of a property varies from one round of a proficiency testing scheme to another, plots of standardized scores, such as *z* and P_A, against the assigned value will show if the participant bias changes with level. When more than one proficiency test item is included in the same round the scores can all be plotted independently.

It can be useful to have a different plotting symbol or different color for the results from the current round of proficiency testing, to distinguish the point(s) from previous rounds.

An example of such a plot is provided in Annex E.10, using P_A scores. This plot could as easily use *z*, with only a change in the vertical scale.

11 Design and analysis of qualitative proficiency testing schemes (including nominal and ordinal properties)

11.1 Types of qualitative data

A large amount of proficiency testing occurs for properties that are measured or identified on qualitative scales. This includes the following:

- Proficiency testing schemes that require reporting on a categorical scale (sometimes called “nominal”), where the property value has no magnitude (such as a type of substance or organism);

- Proficiency testing schemes for presence or absence of a property, whether determined by subjective criteria or by the magnitude of a signal from a measurement procedure. This can be regarded as a special case of a categorical or ordinal scale, with only two values (also called ‘dichotomous’, or binary);

- Proficiency testing schemes requiring results reported on an ordinal scale, which can be ordered according to magnitude but for which no arithmetic relationships exist among different results. For example. “high”, “medium” and “low” form an ordinal scale.

Such proficiency testing schemes require special consideration for the design, value assignment and performance evaluation (scoring) stages because

- Assigned values are very often based on expert opinion

- Statistical treatment designed for continuous-valued and count data is not applicable to qualitative data. For example, it is not meaningful to take means and standard deviations of ordinal scale results even when they can be placed in a ranking order.

The following paragraphs accordingly provide guidance on design, value assignment and performance evaluation for qualitative proficiency testing schemes.

11.2 Statistical design

11.2.1 For proficiency testing schemes in which expert opinion is essential either for value assignment or for assessment of participant reports, it will normally be necessary to assemble a panel of appropriately qualified experts and to provide time for debate in order to achieve consensus on appropriate assignment. Where there is a need to rely on individual experts for scoring or value assignment the proficiency testing provider should additionally provide for assessment and control of the consistency of opinion among different experts.

Example: In a clinical proficiency testing scheme that relies on microscopy for diagnosis, expert opinion is used to assess microscope slides provided to participants and provide an appropriate clinical diagnosis for test items. The proficiency testing provider may choose to circulate test items 'blind' to different members of the expert panel to assure consistency of diagnosis, or carry out periodic exercises to evaluate agreement among the panel.

11.2.2 For proficiency testing schemes that report simple, single-valued categorical or ordinal results, the proficiency testing provider should consider

- providing multiple proficiency test items per round; or
- requesting the results of a number of replicated observations on each proficiency test item, with the number of replicates specified in advance.

Either of these strategies permits counts of results for each participant that can be used either in reviewing data or in scoring. Provision of multiple proficiency test items may provide additional information on the nature of errors and also allow more sophisticated scoring of proficiency testing performance.

Example: In a proficiency testing scheme intended to report the presence or absence of a contaminant, provision of proficiency test items containing a range of levels of the contaminant allows the proficiency testing to examine the number of successful detections at each level as a function of the level of contaminant present. This may be used, for example, to provide information to participants on the detection capability of their chosen test method, or to obtain an average probability of detection which may in turn permit scores to be allocated to participants on the basis of estimated probabilities of particular patterns of response.

11.2.3 Homogeneity should be demonstrated with review of an appropriate sample of proficiency test items, all of which should demonstrate the expected property value. For some qualitative properties, for example presence or absence, it may be possible to verify homogeneity with quantitative measurements; for example a microbiological count or a spectrum absorbance above a threshold. In these situations a conventional test of homogeneity may be appropriate, or a demonstration of all results being above or below a cut-off value.

An example of the analysis of ordinal data is provided in Annex E.11.

11.3 Assigned values for qualitative proficiency testing schemes

11.3.1 Values may be assigned to proficiency test items:
- by expert judgement;
- by use of reference materials as proficiency test items;
- from knowledge of the origin or preparation of the proficiency test item(s);
- using the mode or median of participant results.

Any other value assignment method that can be shown to provide reliable results may also be used. The following paragraphs consider each of the above strategies.

NOTE It is not usually appropriate to provide quantitative information regarding the uncertainty of the assigned value. Each of the paragraphs 11.3.2 to 11.3.5 nonetheless requires the provision of basic information relating to confidence in the assigned value so that participants may judge whether a poor result might reasonably be attributable to an error in value assignment.

11.3.2 Values assigned by expert opinion should normally be based on a consensus of a panel of suitably qualified experts. Any significant disagreement among the panel should be recorded in the report for the round. If the panel cannot reach a consensus for a particular proficiency test item, the proficiency testing provider may consider an alternative method of value assignment from those listed in clause 11.3.1. If that is not appropriate the proficiency test item should not be used for performance assessment of participants.

11.3.3 Where a reference material is provided to participants as a proficiency test item, the associated reference, or certified, value should normally be used as the assigned value for the round. Any summary information provided with the reference material that relates to confidence in the assigned value should be available to participants following the round.

NOTE The limitations of this approach are listed in section 7.4.1.

11.3.4 Where the proficiency test items are prepared from a known source, the assigned value may be determined based on the origin of the constituent(s) of the material. The proficiency testing provider should retain records of the origin, transport and handling of the material(s) used. Due care must be taken to prevent cross-contamination that might result in incorrect results from participants. Evidence of origin and/or detail of preparation should be available to participants after the round either on request or as part of the report for the round.

Example: Proficiency test items of wine circulated for an authenticity proficiency testing scheme may be procured directly from a suitable producer in the designated region of origin, or via a commercial supplier able to provide assurance of authenticity.

11.3.4.1 Confirmatory tests or measurements are recommended where possible, especially where contamination may compromise use as a reference material. For example, a material identified as an exemplar of a single microbial, plant or animal species should normally be tested for response to tests for other relevant species. Such tests should be as sensitive as possible to ensure that contaminating species are either absent or that the level of contamination is quantified.

11.3.4.2 The proficiency testing provider should provide information on any contamination detected that may compromise use of the material. However, it is not usual to provide a quantitative indication of uncertainty of origin on such materials.

NOTE Further detail on characterisation of such materials is beyond the scope of this International Standard.

11.3.5 The mode (the most common observation) may be used as the assigned value for results on a categorical or ordinal scale, while the median may be used as the assigned value for results on an ordinal scale. Where these statistics are used, the report for the round should include a statement of the proportion of the results used in value assignment that matched the assigned value. It is never appropriate to calculate means or standard deviations for proficiency testing results for qualitative properties, including ordinal values. This is because there is no arithmetic relationship between different values on each scale.

11.3.6 When assigned values are based on measurements (for example, presence or absence), the assigned value can usually be determined definitively; i.e., with low uncertainty. Statistical calculations for uncertainty may be appropriate for levels of measurand in "indeterminate" or "equivocal" levels.

11.4 Performance evaluation and scoring for qualitative proficiency testing schemes

11.4.1 Evaluation of participant performance in a qualitative proficiency testing scheme depends in part on the nature of the report required. In some proficiency testing schemes, where a significant amount of evaluation is required of participants and the conclusions require careful consideration and wording, participant reports may be passed to experts for appraisal and may be given an overall mark. At the other extreme, participants may be judged solely on whether their result coincides exactly with the assigned value for the relevant proficiency test item. The following paragraphs accordingly provide guidance on performance assessment and scoring for a range of circumstances.

11.4.2 Expert appraisal of participant reports requires one or more individual experts to review each participant report for each proficiency test item and allocate a performance mark or score. In such a proficiency testing scheme, the proficiency testing provider should ensure that:

- The particular participant is not known to the expert. In particular, the report passed to the expert(s) should not include any information that could reasonably identify the participant;

- review, marking and performance assessment follow a set of previously agreed criteria that are as objective as reasonably possible;

- the provisions of paragraph 11.3.2 with respect to consistency among experts are met;

- where possible, provision is made for participant appeal against a particular expert opinion and/or for secondary review of opinions close to any important performance threshold.

11.4.3 Two systems may be used for scoring a single reported qualitative result based on an assigned value:

i) Each result is marked as acceptable (or scored as a success) if it exactly matches the assigned value and is marked as unacceptable, or given an adverse score, otherwise.

> Example: In a scheme for determining the presence or absence of a contaminant, correct results are scored as 1 and incorrect results as 0.

ii) Results that exactly match the assigned value are marked as acceptable and given a corresponding score; results that do not exactly match the assigned value are given a score that depends on the nature of the mismatch.

> Example 1: In a clinical pathology proficiency testing scheme, a proficiency testing provider assigns a score of '0' for an exactly correct identification of a microbiological species, '1' point for a result that is incorrect but would not change clinical treatment (for example identification as a different but related microbiological species requiring similar treatment), and 3 points for an identification that is incorrect and would lead to incorrect treatment of a patient. Note that this scoring scheme will usually require expert judgement on the nature of the mismatch, perhaps obtained prior to scoring.
>
> Example 2: In a proficiency testing scheme for which six possible responses ranked on an ordinal scale are possible, a result matching the assigned value is given a score of '5' and the score is reduced by 2 for each difference in rank until the score reduces to '0' (so a result adjacent to the assigned value would attract a score of '3').

Individual scores for each proficiency test item should normally be provided to participants. Where replicate observations are performed a summary of scores for that proficiency test item may be provided.

11.4.4 Where multiple replicates are reported for each proficiency test item or where multiple proficiency test items are provided to each participant, the proficiency testing provider may calculate and use combined scores or score summaries in performance assessment. Combined scores or summaries may be calculated as, for example:

- The simple sum of scores across all proficiency test items

- The count of each level of score allocated

- The proportion of correct results

- A distance metric based on the differences between results and assigned values.

> Example: A very general distance metric sometimes used statistics for qualitative data is the Gower coefficient [**18**]. This can combine quantitative and qualitative variables based on a combination of scores for similarity. For categorical or binary data the index allocates a score of 1 for exactly matching categories and 0 otherwise; for ordinal scales it allocates a score equal to 1 minus the difference in rank divided by the number of ranks available, and for interval or ratio scale data it allocates a score equal to 1 minus the absolute difference divided by the observed range of all values. These scores, which are all necessarily from 0 to 1, are summed and the sum divided by the number of variables used. A weighted variant may also be used.

Combined scores may be associated with a summary performance assessment. For example, particular (usually high) proportion of correct scores may be deemed ‘acceptable’ performance, if that is consistent with the objectives of the proficiency testing scheme.

11.4.5 Graphical methods may be used to provide performance information to participants or to provide summary information in a report for a round. An example of the analysis of ordinal data is provided in Annex E.11.

NOTE The above discussion on ordinal data does not apply to measurement results that are based on a quantitative scale with discontinuous indications (such as dilutions, or titres), where the result reflects a magnitude for the quantity and where there is a meaningful arithmetic relationship among the achievable values. In these cases it is usually appropriate to use conventional quantitative statistical techniques, perhaps with some transformation of the results before analysis to meet the assumptions for the distribution (such as by logarithms, for example). Conventional scoring techniques can be used even if the means and standard deviations are not values that can be observed on the discontinuous scale. In some cases the proficiency testing provider may wish to round the evaluation criteria or scores to include achievable values.

Annex A
(normative)

Symbols

g Number of proficiency test items tested in a homogeneity check

m Number of repeat measurements made per proficiency test item

p Number of participants taking part in a round of a proficiency testing scheme

s_s Estimate of between-samples standard deviation

s_x Standard deviation of sample averages

s_w Within sample or within laboratory standard deviation

s_r Estimate of repeatability standard deviation

s_R Estimate of reproducibility standard deviation

s^* Robust estimate of the participant standard deviation

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

$u(x_i)$ Standard uncertainty of a result from participant *i*

$u(x_{ref})$ Standard uncertainty of a reference value

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

$U(x_i)$ Expanded uncertainty of reported result from participant i

$U(x_{ref})$ Expanded uncertainty of a reference value

w_t Between-test-portion range

x Measurement result (generic)

x_i Measurement result from participant *i*

x_{pt} Assigned value

x_{ref} Reference value for a stated purpose

x_{CRM} Assigned value for a property in a Certified Reference Material

x^* Robust average of participant results

$\bar{x}$ Arithmetic average of a set of results

z Score used for proficiency assessment

z' Modified *z* score that includes the uncertainty of the assigned value

ζ Modified *z* score that includes uncertainties for the participant result and the assigned value

E_n “Error, normalized” score that includes uncertainties for the participant result and the assigned value

d Difference between a measurement value for a proficiency test item and an assigned value for a CRM

$\bar{d}$ Average difference between measurement values and the assigned value for a CRM

D Participant difference from the assigned value (x-x_{pt})

$D\%$ Participant difference from the assigned value expressed as a percentage of x_{pt}

δ_E Maximum allowed measurement error criterion for differences

P_A Proportion of allowed error (D/δ_E), can be expressed as a percentage

σ_L Between-laboratory (or participant) standard deviation

σ_{pt} Standard deviation for proficiency assessment

σ_r Repeatability standard deviation

σ_R Reproducibility standard deviation

σ_k bandwidth standard deviation used for kernel density plots

Annex B
(normative)

Homogeneity and stability checks of samples

B.1 General procedure for a homogeneity check

B.1.1 To conduct an assessment for homogeneity for a bulk sample preparation, follow the procedure given below.

a) Choose a property (or properties) to assess with the homogeneity check and its associated measurand(s).

b) Choose a laboratory to carry out the homogeneity check and a measurement method to use. The method should have sufficiently small repeatability so that any significant inhomogeneity can be detected. The ratio of the repeatability standard deviation for the method to the standard deviation for proficiency assessment should be less than 0,5, as recommended in the IUPAC Harmonized Protocol (or 1/6 of δ_E). It is recognized that this is not always possible, so in that case the proficiency testing provider should use more replicates.

c) Prepare and package the proficiency test items for a round of the proficiency testing scheme, ensuring that there are sufficient items for the participants in the scheme and for the homogeneity check.

d) Select a number g of the proficiency test items in their final packaged form using a suitable random selection process, where $g \geq 10$. The number of items included in the homogeneity check may be reduced if suitable data are available from previous homogeneity checks on similar items prepared by the same procedures.

e) Prepare $m \geq 2$ test portions from each proficiency test item using techniques appropriate to the test material to minimize between-test-portion differences.

f) Taking the $g \times m$ test portions in a random order, obtain a measurement result on each, completing the whole series of measurements under repeatability conditions.

g) Calculate the general average $\bar{x}$, within-samples standard deviation s_w, and between-samples standard deviation s_s, as shown in B.3.

B.1.2 When it is not possible to conduct replicate measurements, for example with destructive tests, then the standard deviation of the results can be used as s_s. In this situation it is important to have a method with low repeatability s_r.

B.2 Assessment criteria for a homogeneity check

B.2.1 The following three checks should be used to assure that the homogeneity test data are valid for analysis:

a) Examine the results for each test portion in order of measurement to look for a trend (or drift) in analysis; if there is an apparent trend, then take appropriate corrective action regarding the measurement method, or use caution in the interpretation of the results.

b) Examine the results for sample averages by sample production order; if there is a serious trend that causes the proficiency test item to exceed the criterion at B.2.2 or otherwise prevents use of the proficiency test item, then (i) either assign individual values to each proficiency test item; or (ii) discard a subset of proficiency test items significantly affected and retest the remainder for sufficient homogeneity; or (iii) if the trend affects all proficiency test items, follow the provisions at B.2.4.

c) Compare the difference between replicates (or range, if more than 2 replicates) and, if necessary, test for a statistically significant difference between replicates, using Cochran's test (ISO 5725-2). If the

difference between replicates is large for any pair, review a technical explanation for the difference and if appropriate, remove one or both results from the analysis.

B.2.2 Compare the between-samples standard deviation s_S with the standard deviation for proficiency assessment σ_{pt}. The proficiency test items may be considered to be adequately homogeneous if:

$$s_S \leq 0{,}3\ \sigma_{pt} \tag{B.1}$$

The justification for the factor of 0,3 is that when this criterion is met the between-samples standard deviation contributes no more than about 10 % (0,1 σ_{pt}) of the standard deviation for proficiency assessment (when ±3,0 is the criterion for an action signal).

NOTE equivalently, s_s can be compared to δ_E:

$$s_s \leq 0{,}1\delta_E \tag{B.2}$$

B.2.3 It may be useful to expand the criterion to allow for the actual sampling error and repeatability in the homogeneity check. In these cases, take the following steps:

a) Calculate $\sigma^2_{allow} = (0{,}3\ \sigma_{pt})^2$

b) Calculate $c = F_1\sigma^2_{allow} + F_2 s^2_w$

Where

s_w is the within standard deviation as calculated in B.3 and

F_1 and F_2 are from standard statistical tables, reproduced below in table B.1, for the number of proficiency test items selected and with each item tested in duplicate (from [**29**]))

c) If $s_s > \sqrt{c}$ then there is evidence that the batch of material is not sufficiently homogeneous

Table B.1 — Factors F_1 and F_2 for use in testing for sufficient homogeneity

g	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1.59	1.60	1.62	1.64	1.67	1.69	1.72	1.75	1.79	1.83	1.88	1.94	2.01	2.10
F_2	0.57	0.59	0.62	0.64	0.68	0.71	0.75	0.80	0.86	0.93	1.01	1.11	1.25	1.43

NOTE 1 "g" is the number of proficiency test items selected, tested in duplicate (m=2)

NOTE 2 The two constants in Table B.1 are derived from standard statistical tables as follows:

$F_1 = X^2_{g-1,0.95}/(g-1)$, where $X^2_{g-1,0.95}$ is the value exceeded with probability 0.05 by a chi-squared random variable with $g-1$ degrees of freedom, and
$F_2 = (F_{g-1,g,0.95} - 1)/2$ where $F_{g-1,g,0.95}$ is the value exceeded with probability 0.05 by a random variable with an F-distribution with $g-1$ and g degrees of freedom.

B.2.4 If the criteria in B.2.2 or B.2.3 are not met, the proficiency testing provider shall consider the following actions.

a) Include the between-samples standard deviation in the standard deviation for proficiency assessment, by calculating σ'_{pt} as in equation (B.3). Note this needs to be described fully to participants.

$$\sigma'_{pt} = \sqrt{\sigma^2_{pt} + s^2_S} \tag{B.3}$$

b) Include s_s in the uncertainty of the assigned value and use z' or δ_E' to assess performance;

c) When σ_{pt} is the robust standard deviation of participant results, then the inhomogeneity between proficiency test items is included in σ_{pt} and so the criterion for acceptability of homogeneity can be relaxed, with caution.

If none of a) to c) apply, discard the proficiency test item and repeat the preparation after correcting the cause of inhomogeneity.

B.3 Formulae for homogeneity check

In general designs to test homogeneity, appropriate analysis of variance designs are used to generate estimates for the components of variance, notably the estimate of within-samples standard deviation including measurement repeatability (s_w), and variation between samples (s_s). The technique below is for a chosen number g of proficiency test items, measured in repetition m times.

When the preparation of two or more test portions is not available, for example, for a destructive test and also when it is not feasible to prepare more than one test portion, s_x (see equation B.14) can be used as an alternative to s_s

The data from a homogeneity check are represented by;

$x_{t,k}$

where

t represent the proficiency test item (t = 1,2……..,g)

k represents the test portion (k = 1,2….., m)

Define the proficiency test item average as:

$$\overline{x_t} = \frac{1}{m}\sum_{k=1}^{m} x_k \qquad \text{(B.4)}$$

and the between-test-portion variance as

$$w_t^{\,2} = \frac{1}{(m-1)}\sum_{k=1}^{m}(x_k - \overline{x_t})^2 \qquad \text{(B.5)}$$

Calculate the general average;

$$\overline{\overline{x}} = \frac{1}{g}\sum_{t=1}^{g}\overline{x_t} \qquad \text{(B.6)}$$

the variance of sample averages

$$s_x^{\,2} = \frac{1}{(g-1)}\sum_{t=1}^{g}\left(\overline{x_t} - \overline{\overline{x}}\right)^2 \qquad \text{(B.7)}$$

and the within-samples variance ;

$$s_w^{\,2} = \frac{1}{g}\sum_{t=1}^{g} s_t^{\,2} \qquad \text{(B.8)}$$

Calculate the combined variance of s_s and s_w

$$s_{s,w}^{\,2} = \frac{1}{(g-1)}\sum_{t=1}^{g}\left(\overline{x_t} - \overline{\overline{x}}\right)^2 + (1-\frac{1}{m})s_w^{\,2} = s_s^{\,2} + s_w^{\,2} \qquad \text{(B.9)}$$

Finally, calculate the between-samples variance as

$$s_s^{\,2} = s_{s,w}^{\,2} - s_w^{\,2} = \frac{1}{(g-1)} \sum_{t=1}^{g} \left(\overline{x}_t - \overline{\overline{x}} \right)^2 - \frac{1}{m} s_w^{\,2} \qquad \text{(B.10)}$$

For a common design when *m* is 2, following formulae can be used.

Define the sample averages as:

$$x_{t,\cdot} = (x_{t,1} + x_{t,2}) / 2 \qquad \text{(B.11)}$$

and the between-test-portion ranges as:

$$w_t = |x_{t,1} - x_{t,2}| \qquad \text{(B.12)}$$

Calculate the general average:

$$\bar{x}_{\cdot\cdot} = \sum \bar{x}_{t,\cdot} / g \qquad \text{(B.13)}$$

the standard deviation of sample averages:

$$s_x = \sqrt{\sum (x_{t,\cdot} - \bar{x}_{\cdot,\cdot})^2 / (g-1)} \qquad \text{(B.14)}$$

and the within-samples standard deviation:

$$s_W = \sqrt{\sum w_t^2 / (2g)} \qquad \text{(B.15)}$$

where the summations are over samples (t = 1, 2, ..., g).

Finally, calculate the between-samples standard deviation as:

$$s_S = \sqrt{s_x^2 - (s_W^2 / 2)} \qquad \text{(B.16)}$$

NOTE 1 Between-samples variance s_s^2 often becomes negative when s_s is relatively smaller than s_w. This can be expected when proficiency test items are highly homogeneous.

NOTE 2 Instead of using ranges, one could use between test portion standard deviations such as

$$s_t = w_t / \sqrt{2}$$

An example is provided in Annex E.14

B.4 Procedures for stability checking

B.4.1 Types of stability checks

B.4.1.1 There are two basic types of stability checks.

Stability studies may be required to verify:
- Stability throughout the duration of a proficiency testing scheme round (including or excluding transport)
- Stability for a short time under proposed conditions of transport or storage at the participant's facility.

Where there is reasonable assurance from previous experimental studies, experience, or prior knowledge that instability is unlikely, experimental stability checks may be limited to a check for significant change over the course of the round, carried out during and after the proficiency testing round itself. In other circumstances, studies of transport effects and stability for the typical duration of a round may take the form of planned studies prior to circulation of proficiency test items, either for each round or during early planning and feasibility studies to establish consistent transport and storage conditions. Proficiency testing providers may also wish to check for evidence of instability by checking reported results for a trend with reporting date

B.4.1.2 There are a few considerations important for both types of stability checks:

— All properties that are used in the proficiency testing scheme should be checked or otherwise verified for stability. This can be accomplished with previous experience and technical justification based on knowledge of the matrix (or artefact) and measurand.

— At least 3 proficiency test items should be tested, in duplicate; more samples or more replicates should be used if the repeatability is suspect (e.g., if s_r or $s_w > 0{,}5\sigma_{pt}$).

— The stability measurements should have the same number of replicates as were obtained in the study prior to the commencement of the proficiency testing round (for example, the homogeneity check).

B.4.2 Procedure for checking stability during the course of a proficiency testing round

B.4.2.1 A convenient model for testing stability in proficiency testing is to test a small sample of proficiency test items at the conclusion of a proficiency testing round and compare these with proficiency test items tested prior to the round, to assure that no change occurred through the time of the study. The check may include a check for any effect of transport conditions by additionally exposing the proficiency test items retained for the study duration to conditions representing transport conditions. For studies solely intended to check for transport effects (section B.4.3), the comparison is between proficiency test items that are shipped with proficiency test items that are retained under controlled conditions.

Proficiency testing providers may use the results of homogeneity testing prior to the proficiency testing round instead of selecting and measuring a separate set of proficiency test items.

If a proficiency testing provider includes shipped samples in the stability assessment in B.4.2, then the effects of transport are included in the assessment of stability. If the effects of transport are checked separately, then the procedure described in section B.6 should be used.

This model applies equally to proficiency testing schemes in testing and in calibration.

B.4.2.2 The general procedure for a stability check is as follows:

a) Use the same laboratory, the same measurement method, and measure the same property or properties for the stability check as for the tests prior to the start of the proficiency testing round.

b) Allow a time delay between the homogeneity check and the stability test similar to the time delay that will be experienced by the proficiency test items tested by the participants in the proficiency testing scheme. If feasible, assess proficiency test items under the same transport conditions and at the same times as the round of proficiency testing.

c) Select a number g of the proficiency test items at random, where $g \geq 3$.

d) Make the same number of replicate measurements (m) on all proficiency test items in the stability check.

e) Measure all proficiency test item replicates in random order and under repeatability conditions.

f) Calculate the averages $\bar{y}_1$ and $\bar{y}_2$ of the results for the two groups (before and after) respectively.

B.5 Assessment criterion for a stability check

B.5.1 Compare the general average of the measurements obtained in the homogeneity check (or other prior check) with the general average of the results obtained in the stability check. The samples may be considered to be adequately stable if:

$$|\bar{y}_1 - \bar{y}_2| \leq 0{,}3\sigma_{pt} \quad \text{or} \quad \leq 0{,}1\delta_E \tag{B.17}$$

If it is likely that the intermediate precision of the measurement method (or the uncertainty of measurement of the item) contributed to the inability to meet the criterion, the criterion can be expanded by adding the uncertainty of the difference to σ_{pt} and then checking the difference using the following criterion:

$$|\bar{y}_1 - \bar{y}_2| \leq 0{,}3\sqrt{\sigma_{pt}^2 + 2u(x_i)^2} \tag{B.18}$$

B.5.2 If this criterion is not met, the following options should be considered:

- quantify the effect of instability and take it into account in the evaluation (for example with *z'* scores); or
- examine the sample preparation and storage procedures to see if improvements are possible; or
- do not evaluate participant performance.

B.6 Stability in transport conditions

B.6.1 The proficiency testing provider should check the effects of transport of proficiency testing items. Ideally, proficiency test items checked for stability will be shipped at the same time and manner as other proficiency test items, but be returned to the proficiency testing provider.

B.6.2 Any known effects of transportation should be considered when evaluating performance. Any increase in uncertainty due to transport should be included in the uncertainty of the assigned value.

B.6.3 If the assigned value and standard deviation for proficiency assessment are determined from participant results (e.g., by robust methods), then the average and the standard deviation for proficiency assessment might reflect any bias and the uncertainty (respectfully) caused by transport conditions.

B.6.4 The criterion for sufficient stability due to transport is the same as in section B.5.

An example of a stability check is shown in section E.13

Annex C
(normative)

Robust analysis

C.1 Robust analysis: Introduction

Interlaboratory comparisons present unique challenges for data analysis. The expectation of normally distributed (or at least unimodal and symmetric) data for the majority of participants is usually met, but the data set usually includes a significant proportion of results that may not come from competent participants, or come from participants who did not understand the instructions or processed the proficiency test items improperly. These results can be outliers or can create skewness that make it difficult to calculate summary statistics with conventional statistical techniques.

It is recommended that proficiency testing providers use statistical techniques that are robust to the outliers and skewness. Many such techniques have been proposed in the statistical literature, and many of those have been used successfully for proficiency testing. This Annex describes four techniques that have been widely applied and have different capabilities regarding robustness to contaminated populations (for example, efficiency and breakdown point), and regarding simplicity of application. They are presented here in order of simplicity (simplest first, most complex last), which is usually inversely related to efficiency and breakdown point.

NOTE Robustness is a property of the estimation algorithm, not of the estimates it produces, so it is not strictly correct to call the averages and standard deviations calculated by such an algorithm "robust". However, to avoid the use of excessively cumbersome terminology, the terms "robust average" and "robust standard deviation" should be understood in this International Standard to mean estimates of the population mean or of the population standard deviation calculated using a robust algorithm.

Further information on these techniques and special considerations are presented in the informative Annex D.

C.2 Simple robust estimates for the population mean and standard deviation

In some circumstances it could be appropriate to use a simplified robust estimate of the mean and standard deviation. The population median is an acceptable robust estimate of the mean, derived as x^* in Algorithm A, formula (C.2). Similarly, the statistic s^* in formula (C.3) can be an acceptable robust estimate of the population standard deviation; this statistic is based on the 'Median Absolute Deviation' (or MAD), and is a 'consistent' estimate of the population standard deviation for a normal (or Gaussian) distribution. See Note 2 in section C.3 for the situation when 50% or more of the participant results are the same as the median.

A very similar robust estimate of the standard deviation has proved to be useful in many proficiency testing schemes, and can be obtained from the difference between the 75th percentile (or 3rd quartile) and 25th percentile (or 1st quartile) of the participant results, when they are ranked in non-decreasing order (that is, ascending order with all ties listed). This statistic is commonly called the 'normalized InterQuartile Range' (or *nIQR*), and it is calculated as in formula (C.1).

$$nIQR = 0{,}7413(Q_3(x_i) - Q_1(x_i)) \qquad \text{(C.1)}$$

where

$Q_1(x_i)$ = 25th percentile of x_i ($i=1,2,\ldots,p$)

$Q_3(x_i)$ = 75th percentile of x_i ($i=1,2,\ldots,p$)

If the 75th and 25th percentiles are the same, the *nIQR* will be zero and procedure C.2 (including Note 2) should be used to calculate the robust standard deviation s^*.

C.3 Robust analysis: Algorithm A

This algorithm yields robust estimates of the mean and standard deviation of the data to which it is applied.

NOTE 1 Algorithms A and S given in this annex are reproduced from ISO 5725-5.

Denote the *p* items of data, sorted into increasing order, by:

$x_1, x_2, ..., x_p$

Denote the robust average and robust standard deviation of these data by x^* and s^*.

Calculate initial values for x^* and s^* as:

x^* = median of x_i (i = 1, 2, ..., p) (C.2)

s^* = 1,483 median of $|x_i - x^*|$ with (i = 1, 2, ..., p) (C.3)

NOTE 2 In some cases more than half of the results x_i will be the identical (for example, thread count in fabric, or electrolytes in serum). In these cases the initial value of s^* will be zero and the robust procedure will not perform correctly. In the case that the initial $s^* = 0$, it is acceptable to substitute the sample standard deviation, after checking for any gross outliers that could make the sample standard deviation unreasonably large This substitution is made only for the initial s^*, and after that the iterative algorithm can proceed as described.

Update the values of x^* and s^* as follows. Calculate:

$$\delta = 1{,}5s^* \tag{C.4}$$

For each x_i (i = 1, 2, ..., p), calculate:

$$x_i^* = \begin{cases} x^* - \delta, & \text{when } x_i < x^* - \delta \\ x^* + \delta, & \text{when } x_i > x^* + \delta \\ x_i \ , & \text{otherwise} \end{cases} \tag{C.5}$$

Calculate the new values of x^* and s^* from:

$$x^* = \sum x_i^* / p \tag{C.6}$$

$$s^* = 1{,}134\sqrt{\sum (x_i^* - x^*)^2/(p-1)} \tag{C.7}$$

where the summation is over i.

The robust estimates x^* and s^* may be derived by an iterative calculation, i.e. by updating the values of x^* and s^* several times using the modified data in equations C.4 to C.7, until the process converges. Convergence may be assumed when there is no change from one iteration to the next in the third significant figures of the robust standard deviation and the robust average. Alternative convergence criteria can be determined according to the design and reporting requirements for proficiency test results.

C.4 Algorithm S

This algorithm is applied to standard deviations (or ranges), which are calculated when participants submit *m* replicate results for a measurand in a proficiency test item, or in a study with *m* identical proficiency test items. It yields a robust pooled value of the standard deviations or ranges to which it is applied.

Denote the *p* items of data, sorted into increasing order, by:

$w_1, w_2, ..., w_p$

Denote the robust pooled value by w^*, and the degrees of freedom associated with each w_i by ν. (When w_i is a range, $\nu = 1$. When w_i is the standard deviation of m test results, $\nu = m - 1$.) Obtain the values of ξ and η required by the algorithm from Table C.1.

Calculate an initial value for w^* as:

$$w^* = \text{median of } w_i \quad (i = 1, 2, ..., p) \tag{C.8}$$

NOTE If more than half of the w_i are zero then the initial w^* will be zero and the robust procedure will not perform correctly. When the initial w^* is zero, then substitute the arithmetic pooled average standard deviation (or average range) after eliminating any extreme outliers that can influence the average. This substitution is only for the initial w^*, after which the procedure should continue as described.

Update the value of w^* as follows. Calculate:

$$\psi = \eta \times w^* \tag{C.9}$$

For each w_i ($i = 1, 2, ..., p$), calculate:

$$w_i^* = \begin{cases} \psi \text{ , if } w_i > \psi \\ w_i \text{ , otherwise} \end{cases} \tag{C.10}$$

Calculate the new value of w^* from:

$$w^* = \xi\sqrt{\sum (w_i^*)^2/p} \tag{C.11}$$

The robust estimate w^* is calculated by an iterative calculation by updating the value of w^* several times, until the process converges. Convergence may be assumed when there is no change from one iteration to the next in the third significant figure of the robust estimate.

Table C.1 — Factors required for robust analysis: Algorithm S

Degrees of freedom ν	Limit factor η	Adjustment factor ξ
1	1,645	1,097
2	1,517	1,054
3	1,444	1,039
4	1,395	1,032
5	1,359	1,027
6	1,332	1,024
7	1,310	1,021
8	1,292	1,019
9	1,277	1,018
10	1,264	1,017
NOTE The values of ξ and η are derived in Annex B of ISO 5725-5:1998.		

C.5 Computationally intense robust estimates for population mean and standard deviation

C.5.1 Introduction to the Q method

The robust estimators of the population mean and standard deviation described in sections C.2 and C.3 are useful when computational resources are limited, or when it is necessary to provide concise explanations of the statistical procedures. These procedures have proven to be useful in a wide variety of situations, including for proficiency testing schemes in new areas of testing or calibration and in economies where proficiency testing has not previously been available. However, these techniques can break down (i.e., provide unreliable estimates)

— when more than 20% of results are outliers;
— when there are a considerable number of results below the limit of quantitation;
— when there are bimodal (or multimodal) distributions;
— when the underlying distribution of data is highly skewed; or
— when many test results are equal, due to quantitative data on a discontinuous scale or due to rounding distortions.

ISO/IEC 17043 requires that these situations will be anticipated by design or will be detected by competent review prior to performance evaluation, but there are occasions when this may not be possible.

In addition, some of the robust techniques described in sections C.2 and C.3 are lacking in terms of statistical efficiency - if the number of participants is less than 50, there is a considerable risk for misclassifying participants due to the use of ineffective statistical methods.

Robust techniques addressing the above mentioned issues require more intense computational resources and complicated explanations, but the techniques are referenced in available literature and International Standards.

In the following sections a high-efficiency, high-breakdown method is described that has been adapted to proficiency testing requirements and has proven to be useful in a wide variety of situations.

C.5.2 Determination of robust estimate of the standard deviation, s^*, using the Q method

The Q method produces a robust estimate of the standard deviation of proficiency testing results reported by different laboratories. It can be used with and without replicate measurement results. The Q method can be used for proficiency testing both with and without measurement repetitions.

Denote the reported measurement results, grouped by laboratory, by:

$$\underbrace{y_{11}, \dots, y_{1n_1}}_{Lab\ 1}, \underbrace{y_{21}, \dots, y_{2n_2}}_{Lab\ 2}, \ \dots, \underbrace{y_{p1}, \dots, y_{pn_p}}_{Lab\ p}$$

Calculate the cumulative distribution function of all absolute between-laboratory differences

$$H_1(x) = \frac{1}{\binom{p}{2}} \sum_{1 \le i < j \le p} \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{m=1}^{n_j} 1\{|y_{ik} - y_{jm}| \le x\} \qquad \text{(C.12)}$$

Denote the discontinuity points of $H_1(x)$ by:

$x_1, \dots, x_r$, where $x_1 < x_2 < \cdots < x_r$.

Calculate for all positive discontinuity points $x_1, \dots, x_r$:

$$G_1(x_i) = \begin{cases} 0.5 \cdot \left(H_1(x_i) + H_1(x_{i-1})\right) & \text{if} \quad i \ge 2 \\ 0.5 \cdot H_1(x_1) & \text{if} \quad i = 1 \text{ and } x_1 > 0 \end{cases} \qquad \text{(C.13)}$$

and let

$G_1(0) = 0.$

Calculate the function $G_1(x)$ for all x out of the interval $[0, x_r]$ by linear interpolation between discontinuity points $0 \leq x_1 < x_2 < \cdots < x_r$.

Calculate the robust standard deviation s^* of test results of different laboratories

$$s^* = \frac{G_1^{-1}\left(0.25+0.75 \cdot H_1(0)\right)}{\sqrt{2}\Phi^{-1}\left(0.625+0.375 \cdot H_1(0)\right)}. \qquad \text{(C.14)}$$

NOTE 1 This algorithm does not depend on a mean value; it can be used together with both a value from combined participant results (preferably the Hampel estimator described in the paragraph below) or a specified reference value.

NOTE 2 The Q method is a robust method of determining both the standard deviation between laboratories (reproducibility standard deviation) and the standard deviation within laboratories (between multiple determinations), see [24,32]. It is fully characterized and validated by means of asymptotic theory and finite sample breakdown point [**23**].

NOTE 3 No iteration is required.

NOTE 4 Measurement results below the limit of quantitation (LOQ) should not be excluded from the dataset but may be replaced by LOQ/2. As long as the percentage of these results is not higher than 25 – 40 % (depending on the number of laboratories and on the agreement of the remaining data), the corresponding standard deviation obtained by the Q method can be considered reliable. [**24**]

NOTE 5 Annex E presents an example to explain the estimation principle of the Q method.

C.5.3 Determination of robust mean by the Hampel estimator

This algorithm yields a robust value of the mean of the test results reported by different laboratories.

Calculate the arithmetic means for each laboratory, now labelled $y_1, y_2, \ldots y_p$.

Calculate the robust mean, x^*, by solving the equation:

$$\sum_{i=1}^{p} \Psi\left(\frac{y_i - x^*}{s^*}\right) \qquad \text{(C.15)}$$

where

$$\Psi(q) = \begin{cases} 0 & q \leq 4{,}5 \\ -4{,}5 - q & -4{,}5 < q \leq -3 \\ -1{,}5 & -3 < q \leq -1{,}5 \\ q & -1{,}5 < q \leq 1{,}5 \\ 1{,}5 & 1{,}5 < q \leq 3 \\ 4{,}5 - q & 3 < q \leq 4{,}5 \\ 0 & q > 4{,}5 \end{cases} \qquad \text{(C.16)}$$

and s^* is the robust reproducibility standard deviation according to the Q method.

The exact solution may be obtained in a finite number of steps, which means not iteratively, using the property that Ψ in the argument of x^* is partially linear, bearing in mind that the interpolation nodes on the left side of equation (C.16) (interpreted here as a function of x^*) are as follows:

$y_j + d \cdot s^*$ with d = – 4,5; – 3; –1,5; 0; 1,5; 3 and 4,5.

The solution nearest the median is used, but if this does not yield a clear result, the median itself is used as location parameter.

NOTE If this estimation method is used, laboratory results differing from the mean by more than 4,5 times thereproducibility standard deviation no longer have any effect on the calculation result, i.e. they are treated as outliers.

Annex D
(Informative)

Additional Guidance on Statistical Procedures

D.1 Procedures for small numbers of participants

D.1.1 General considerations

Many proficiency testing schemes have few participants, or have comparison groups with small numbers of participants, even if there are a large number of participants in the scheme. This can happen frequently when participants are grouped and scored by method, as is commonly done in proficiency testing for medical laboratories, for example.

Where the number of participants is small, the assigned value should ideally be determined using a metrologically valid procedure, independent of the participants, such as by formulation or from a reference laboratory. Performance evaluation criteria should also be based on external criteria, such as expert judgement or criteria based on fitness for purpose. In these ideal situations, proficiency testing can be conducted with just one participant, and performance is determined by the pre-determined assigned value and performance criterion.

Where these ideal conditions cannot be met, either the assigned value or the dispersion, or both, may need to be derived from participant results. If the number of participants is too small for the particular procedures used the performance evaluation may become unreliable; it is therefore important to consider whether a minimum number of participants should be set for performance evaluation.

The following paragraphs present guidance for situations of small numbers, when the performance evaluation criteria are not determined independently.

NOTE ISO/IEC 17043 states that the proficiency testing provider must fully describe the statistical methods used, they must be fit for purpose, and they must be statistically valid. Furthermore, any statistical assumptions on which the methods are based must be stated in the design, and these assumptions shall be demonstrated to be reasonable.

D.1.2 Procedures for identifying outliers

Although robust statistics are strongly recommended for outlier-contaminated populations, they are not often recommended for very small data sets (see below for exceptions). Outlier testing, however, is possible for very small data sets. Outlier rejection followed by, for example, calculation of the mean or standard deviation may therefore be preferable in the case of very small schemes or groups.

Different outlier tests are applicable to different data set sizes. ISO 5725-2 provides tables for Grubbs test 1 and Dixon's test for single outlier tests for outlying participant means when $p \geq 3$. Tables are provided for Grubbs tests 2 and 3, which test for two simultaneous outliers. Both Grubbs and Dixon's tests require the number of possible outliers to be specified in advance and can fail when there are multiple outliers, making them most useful for $p < 10$ (depending on the likely proportion of outliers).

Note 1 Care should be taken when estimating dispersion after outlier rejection as dispersion estimates will be biased low. The bias is not usually serious if rejection is carried out only at the 99% level of confidence or above.

Note 2 Most univariate robust estimators for location and scale perform acceptably for $p \geq 15$.

D.1.3 Procedures for estimates of location

D.1.3.1 Assigned values derived from small sets of participant data should, where possible, meet the criterion for uncertainty of the assigned value given at 9.2.1. For a simple mean and a standard deviation of results equal to the standard deviation for proficiency assessment, this criterion cannot be met for a normal distribution with $p \leq 12$, after any removal of outliers. For the median (taking the efficiency as 0.64), the criterion cannot be met for $p \leq 18$. Other robust estimators, such as Algorithm A (C.3), have intermediate efficiency and may meet the criterion for $p > 12$ if the provisions of 7.7.2.2 NOTE 2 are taken into account.

D.1.3.2 There are data set size limitations on the applicability of some estimators of location. Few computationally intensive robust estimators for the mean are recommended for small data sets; a typical lower limit is $p \geq 15$, though providers may be able to demonstrate acceptable performance for specific assumptions on smaller data sets. The median is applicable down to $p = 2$ (when it is equal to the mean) but at $3 \leq p \leq 5$ the median offers few advantages over the mean unless there is an unusually high risk of poor results.

D.1.4 Procedures for estimates of dispersion

D.1.4.1 Use of performance criteria based on the dispersion of participant results is not recommended for small data sets owing to the very high variability of any dispersion estimates. For example, for $p = 30$ estimates of the standard deviation for normally distributed data are expected to vary by approximately 25% either side of its true value (based on a 95% confidence level). No other estimator improves on this for normally distributed data.

D.1.4.2 Where dispersion estimators are required for other purposes (for example as summary statistics or an estimate of scale for robust location estimators), or where the scheme can tolerate high variability in dispersion estimates, dispersion estimates with the highest available efficiency should be selected when handling smaller data sets.

Note 1 'Highest available' is understood to take account of availability of suitable software and expertise.

Note 2 The Q_n estimate of scale described in ISO 16269-4 [**10**] is considerably more efficient than either the median absolute deviation (MAD) or *nIQR*.

Note 3 Specific recommendations have been made for robust estimates of dispersion in very small data sets [**24**] as follows:

- $p \leq 2$: use the mean.
- $p = 3$, locations and scale unknown: use scaled median absolute deviation to protect against excessively high scale estimates or the mean absolute deviation to protect against unduly small scale estimates (for example when rounding may give two identical values)
- $p \geq 4$: A specific *M*-estimate of scale based on a logarithmic weighting function was recommended by reference [**24**]; a near equivalent is Algorithm A with no iteration of location, using the median as a location estimate.

Note 4 To obtain an estimate of standard deviation from the absolute distance to the median, use

$$s^* = \frac{1}{0.798\ p} \sum_{i=1,\ p} \left| x_i - \text{median}\ (x) \right|$$

D.2 Efficiency and breakdown points for robust procedures

D.2.1 Different statistical estimators (e.g., robust techniques) can be compared on two key characteristics:

Breakdown point – the proportion of outliers in a dataset that can lead to incorrect estimates.

Efficiency – the ratio of the theoretical minimal variance to the actual variance of an estimator.

These characteristics depend heavily on the underlying distribution of results for a population of competent participants, and the nature of results that are from incompetent participants (or from participants that did not follow instructions or the measurement method). The contaminated data can appear as outliers, results with larger variance, or results with a different mean (e.g., bimodal).

Breakdown points and efficiencies for the different estimators will be different for different situations, and a thorough review is beyond the scope of this document. However simple comparisons can be made under the

assumption of a normal distribution for results from competent laboratories, with a mean equal to x_{pt} and standard deviation equal to σ_{pt}.

D.2.2 Breakdown point

Breakdown point is the limit for the proportion of results that can be outliers without affecting the statistical estimator. It should be noted that procedures required in sections 6.3 and 6.4 should prevent data analysis of datasets with large proportions of outliers. However there are situations where visual review is not practical.

Table D.1 — Breakdown points for estimates of the mean and standard deviation (proportion of outliers that can lead to failure of the estimator)

Statistical estimator	Population parameter estimated	Breakdown Point
Median	Mean	50%
nIQR	Standard Deviation	25%
MAD	Standard Deviation	50%
Algorithm A	Mean and SD	25%
Q method/Hampel	Mean and SD	50%
Arithmetic mean	Mean and SD	≈ 0-10 %
Classical standard deviation	Standard Deviation	≈ 0-10 %

In summary, the classic mean and standard deviation can break down with only a single outlier, although outlier removal could lead to better resistance to outliers, if the outliers are large enough. This is not always the case. The robust methods using the median, MAD, and Q/Hampel methods can tolerate a very large proportion of outliers because, they are based largely on the midpoints of the distribution. The robust methods nIQR and Algorithm A can break down with 25% of data as outliers (if they are on the same side of the median), because they are based on the middle 50% of results – the values between the 25th and 75th percentile of results (1st and 3rd quartiles).

D.2.3 Statistical efficiency

All estimators have sampling variance – that is, the estimators can vary from study to study, even if all participants are competent and there are no outliers or subgroups of participants with different means or variances. Robust estimators modify submitted results that are exceptionally far from the middle of the distribution, based on theoretical assumptions, and so these estimators have larger variance than the minimum variance estimators, in the case that the dataset is in fact normally distributed.

The classic arithmetic mean and standard deviation are the minimum variance estimators of the population mean and standard deviation, and so they have efficiency of 100%. Estimators with lower efficiency have higher variance – that is, they could vary more from study to study, even if there are no outliers or different subgroups of participants.

Table D.2 — Efficiency of robust estimates for the population mean and standard deviation, for datasets with n=50 or 500 participants:

Estimator	Mean, n=50	Mean, n=500	SD, n=50	SD, n=500
Median	66%	65%	NA	NA

nlQR	NA	NA	38%	37%
MAD	NA	NA	37%	37%
Algorithm A	97%	97%	74%	73%
Q method/Hampel	96%	96%	73%	81%
Classical mean & SD	100%	100%	NA	NA

These results demonstrate that there is no statistical method that is perfect for all situations. Arithmetic mean and classical standard deviation are optimal with a normal distribution but break down in case of outliers. Simple robust methods such as median, MAD or nlQR perform well only in presence of outliers. They should only be used if no other algorithm is available.

D.3 Use of proficiency testing data for monitoring and estimating measurement uncertainty

D.3.1 When assigned values are appropriate for a given application, participant results should be close to the assigned values, within the claimed expanded uncertainties for the results. The differences between participant results and the assigned values can be monitored across measurand levels and different rounds of proficiency testing to verify the appropriateness of uncertainty claims. *Zeta* scores can be useful for this, and participants should calculate their own *zeta* scores if the proficiency testing provider does not do so. Assessment and monitoring of uncertainty can be accomplished only when the assigned values are metrologically traceable to an appropriate reference.

D.3.2 In some situations, it can be possible for a participant to estimate the uncertainty of measurement for their results, using data from multiple rounds of proficiency testing (see ISO 11352:2012 [**7**]), using the following equation:

$$u_c = \sqrt{u_{R_W}^2 + u_b^2} \tag{19}$$

Where

u_c = combined standard uncertainty for the participant's results

u_{R_W} = standard uncertainty associated with within-laboratory variability

u_b = standard uncertainty associated with method and participant effects

NOTE 1 the component u_b is derived from the variation between differences or relative differences, D *or* $D\%$, from multiple rounds of proficiency testing, and the uncertainty of the assigned values.

NOTE 2 the component u_{R_W} is determined in different ways for different conditions, and it is intended to reflect within-laboratory variation over time, which is usually larger than repeatability (or short term variability).

NOTE 3 ISO 11352:2012 should be referenced for more a complete description of these components.

D.3.3 In order for the resulting estimate of uncertainty to be valid, the following should apply:

a) The assigned values for the different rounds shall be metrologically traceable to the same reference;
b) The participant's uncertainty shall be assumed to be constant across the different levels evaluated;
The participant's method shall be the same for all rounds evaluated.

D.3.4 If these conditions do not apply or if the series of proficiency testing schemes extends longer than 2 years, this technique should not be used.

D.4 Use of proficiency testing data for evaluating the reproducibility and repeatability of a measurement method

D.4.1 The Introduction to ISO/IEC 17043 states that the evaluation of the performance characteristics of a method is generally not a purpose of proficiency testing. However, it can be possible to use the results of proficiency testing schemes to verify, and perhaps establish the repeatability and reproducibility of a measurement method when the scheme meets the following conditions:

a) several rounds of the scheme have demonstrated that samples are sufficiently homogeneous and stable;
b) participants are capable of consistent satisfactory performance,
c) the competence of participants (or a subset of participants) has been demonstrated prior to the round, and their competence is not placed in doubt by the results in the round.

D.4.2 In order to accomplish this, the following design conditions must be used:

a) There are at least 8 participants that have demonstrated competence with a measurement method on previous rounds, and have committed to follow the measurement method without modification;
b) The round includes at least 2 duplicate samples;
c) The participants do not know that the samples are duplicates;
d) Samples used in one or several rounds of the scheme cover the range of levels and types of samples for which the measurement method is intended;
e) Data analysis procedures are consistent with ISO 5725-2 for calculating repeatability and reproducibility

Annex E
(Informative)

Illustrative Examples

These examples are intended to illustrate the procedures specified in this Standard, so the reader can determine that their calculations are correct.

E.1 Effect of censored values Section 5.5.3

Table E.1 shows 23 results for a PT exercise, of which 5 results are indicated as 'Less Than' some amount. The robust mean (x*) and standard deviation (s*) are shown for 3 different calculations, where the '<' signs are discarded and data analysed as quantitative data; the results with '<' values are ignored; and where 0.5 times the result is inserted as an estimate of the quantitative result. In each scenario the results that would have been outside the acceptance limit have been highlighted. This assumes that the evaluation would be 'unacceptable' for any result where the quantitative part is outside the x* +/- s*. The proficiency testing provider could have alternative rules for evaluating results with '<' or '>' signs.

Table E.1 — Sample dataset with truncated (<) results, and three options for accommodating results.

Participant	Result	'<' ignored	'< ' deleted	0.5 * '<'
A	<10	10		5
B	<10	10		5
C	12	12	12	12
D	19	19	19	19
E	<20	20		10
F	20	20	20	20
G	23	23	23	23
H	23	23	23	23
J	25	25	25	25
K	25	25	25	25
L	26	26	26	26
M	28	28	28	28
N	28	28	28	28
P	<30	30		15
Q	28	28	28	28
R	29	29	29	29
S	30	30	30	30
T	30	30	30	30
U	31	31	31	31
V	32	32	32	32
W	32	32	32	32
Y	45	45	45	45
Z	<50	50		25
Results	23	23	18	23

x*		26,02	26,81	23,98
s*		7,11	5,29	8,52

E.2 Comprehensive example of Atrazine in Drinking Water (courtesy, Univ. Stuttgart)

A study of Atrazine (a herbicide) in drinking water. This PT is graded in various ways by different regulatory authorities:
Using the robust mean of participant results as the assigned value
Using the gravimetric value for the pure weighed-in amount of atrazine into clean water, as assigned value
Using the robust standard deviation of results as the standard deviation for proficiency assessment.
Using a fit-for-purpose allowance of total error from the gravimetric assigned value.
Using participant-submitted uncertainties to calculate zeta scores.

Participants were asked to submit their expanded uncertainty estimates and their coverage factor; these results were transformed by dividing the participant's expanded uncertainty by the coverage factor, to calculate a combined standard uncertainty. This value is compared with the uncertainty of the assigned value (u_{min}) and 1.5 times the robust standard deviation (u_{max}).

This example shows the raw data as submitted (but ordered by value for clarity). It shows calculated values for the robust mean and standard deviation following Algorithm A; it also shows other calculated values for the robust estimate of the mean and standard deviation.
The tables show the various summary statistics and the resulting performance scores, calculated with the assigned values and performance scores indicated.

Table E.2a — Calculation of the robust average and standard deviation for Atrazine in drinking water

Iteration	0	1	2	3	4	5
δ = 1,5 s*	—	0.0578	0.0581	0.0587	0.0590	0.0592
$x^* - \delta$	—	0.2042	0.1997	0.1985	0.1980	0.1979
$x^* + \delta$	—	0.3198	0.3160	0.3159	0.3161	0.3162
78	0.040	0.204	0.200	0.198	0.198	0.198
22	0.055	0.204	0.200	0.198	0.198	0.198
114	0.178	0.204	0.200	0.198	0.198	0.198
42	0.202	0.204	0.202	0.202	0.202	0.202
106	0.206	0.206	0.206	0.206	0.206	0.206
93	0.227	0.227	0.227	0.227	0.227	0.227
10	0.228	0.228	0.228	0.228	0.228	0.228
19	0.230	0.230	0.230	0.230	0.230	0.230
26	0.230	0.230	0.230	0.230	0.230	0.230
50	0.235	0.235	0.235	0.235	0.235	0.235
100	0.236	0.236	0.236	0.236	0.236	0.236
39	0.237	0.237	0.237	0.237	0.237	0.237
20	0.243	0.243	0.243	0.243	0.243	0.243
45	0.244	0.244	0.244	0.244	0.244	0.244
32	0.245	0.245	0.245	0.245	0.245	0.245
33	0.2555	0.256	0.256	0.256	0.256	0.256
14	0.260	0.260	0.260	0.260	0.260	0.260
27	0.264	0.264	0.264	0.264	0.264	0.264

91	0.267	0.267	0.267	0.267	0.267	0.267
79	0.270	0.270	0.270	0.270	0.270	0.270
30	0.273	0.273	0.273	0.273	0.273	0.273
4	0.274	0.274	0.274	0.274	0.274	0.274
24	0.274	0.274	0.274	0.274	0.274	0.274
110	0.278	0.278	0.278	0.278	0.278	0.278
37	0.2811	0.281	0.281	0.281	0.281	0.281
105	0.287	0.287	0.287	0.287	0.287	0.287
108	0.287	0.287	0.287	0.287	0.287	0.287
43	0.288	0.288	0.288	0.288	0.288	0.288
75	0.289	0.289	0.289	0.289	0.289	0.289
102	0.295	0.295	0.295	0.295	0.295	0.295
99	0.296	0.296	0.296	0.296	0.296	0.296
40	0.311	0.311	0.311	0.311	0.311	0.311
11	0.331	0.320	0.316	0.316	0.316	0.316
68	0.4246	0.320	0.316	0.316	0.316	0.316
Average	0.2512	0.2579	0.2572	0.2571	0.2570	0.2570
Standard deviation	0.0672	0.0342	0.0345	0.0347	0.0348	0.0348
New *x**	**0.2620**	**0.2578**	**0.2572**	**0.2571**	**0.2570**	**0.2570**
New *s**	**0.0386**	**0.0387**	**0.0391**	**0.0393**	**0.0394**	**0.0395**

Table E.2b — Summary Statistics for Atrazine example

Procedure	Estimate of Mean	Estimate of SD	Uncertainty of Mean
Robust: Median, nIQR (MAD)	0,2620	0,0402 (0,0386)	0,0084
Robust: Algorithm A (x*, s*)	0,2570	0,0395	0,0082
Robust: Hampel Q	0,2600	0,0426	0,0089
Reference values	0,2753	0,0413 (15%)	0,0046 (0,0018 for gravimetric)
Arithmetic, outliers removed	0,2588	0,0337	0,0061
Arithmetic, outliers included (all data)	0,2512	0,0672	nc

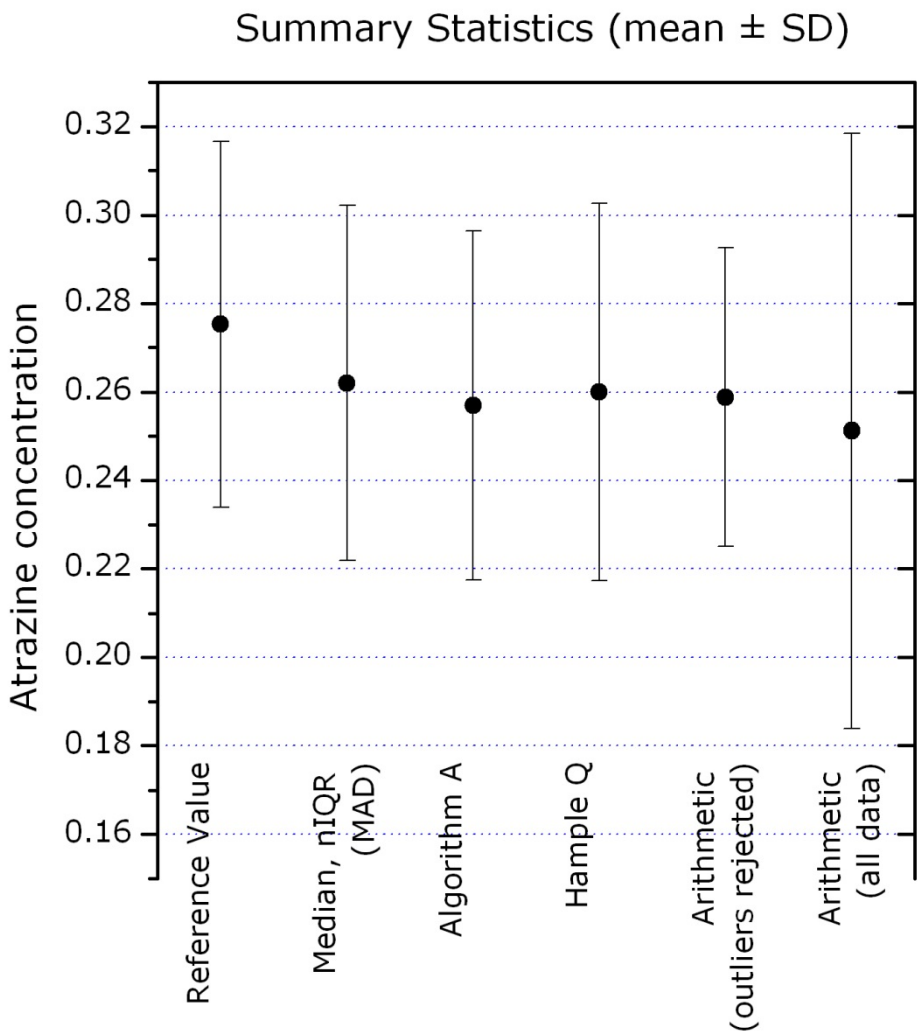


Figure E.2a — Summary of Mean ± 1SD for various calculation methods

Table E.2c — Performance statistics for assigned value = 0,2753 (Reference value) - σ_{pt} is SD Hampel Q (0,0426) || u_X from reference =0,0018 || and δ_E = 45% of assigned value

Lab/Statistic	D%	PA	z	z'	zeta
78	-85.5%	-189.9%	-5.53	-5.52	-2.3
22	-80.0%	-177.8%	-5.18	-5.17	-10.5
114	-35.3%	-78.5%	-2.29	-2.28	-3.6
42	-26.6%	-59.2%	-1.72	-1.72	-8.0
106	-25.2%	-55.9%	-1.63	-1.63	-38.5
93	-17.5%	-39.0%	-1.13	-1.13	-0.3
10	-17.2%	-38.2%	-1.11	-1.11	-1.5
19	-16.5%	-36.6%	-1.06	-1.06	-25.2

26	-16.5%	-36.6%	-1.06	-1.06	-2.2
50	-14.6%	-32.5%	-0.95	-0.95	-22.4
100	-14.3%	-31.7%	-0.92	-0.92	-19.1
39	-13.9%	-30.9%	-0.90	-0.90	-2.5
20	-11.7%	-26.1%	-0.76	-0.76	-17.9
45	-11.4%	-25.3%	-0.74	-0.73	-1.5
32	-11.0%	-24.5%	-0.71	-0.71	-16.8
33	-7.2%	-16.0%	-0.47	-0.46	-0.5
14	-5.6%	-12.4%	-0.36	-0.36	-0.6
27	-4.1%	-9.1%	-0.27	-0.27	-0.6
91	-3.0%	-6.7%	-0.19	-0.19	-4.6
79	-1.9%	-4.3%	-0.12	-0.12	-0.3
30	-0.8%	-1.9%	-0.05	-0.05	-0.2
4	-0.5%	-1.0%	-0.03	-0.03	-0.7
24	-0.5%	-1.0%	-0.03	-0.03	-0.7
110	1.0%	2.2%	0.06	0.06	1.0
37	2.1%	4.7%	0.14	0.14	3.2
105	4.2%	9.4%	0.27	0.27	0.1
108	4.2%	9.4%	0.27	0.27	6.5
43	4.6%	10.3%	0.30	0.30	1.2
75	5.0%	11.1%	0.32	0.32	0.9
102	7.2%	15.9%	0.46	0.46	10.9
99	7.5%	16.7%	0.49	0.49	0.5
40	13.0%	28.8%	0.84	0.84	19.8
11	20.2%	45.0%	1.31	1.31	1.7
68	54.2%	120.5%	3.51	3.50	3.0

Table E.2d — PT Results and Uncertainties for participant results u_{min} = 0,0018 u_{max} = 1,5*0,0426 with flags for u_{lab}<u_{min} = *a* or u_{lab}>u_{max} = *c* and with indicator for whether a result was an outlier by Grubbs test (α=0,05) replicated

Lab	Result	U_{lab}	k	u_{lab}	U flag	Grubbs
78	0.040	0.201	2	0.1005	b	Y
22	0.055	0.041	1.96	0.0209	b	Y
114	0.178	0.054	2	0.027	b	
42	0.202	0.018	2	0.009	b	
106	0.206	0			a	
93	0.227	0.276	2	0.138	c	
10	0.228	0.065	2	0.0325	b	
19	0.230	0			a	
26	0.230	0.036	1.732	0.0208	b	
50	0.235	0			a	
100	0.236	0.002	2	0.001	a	
39	0.237	0.031	2	0.0155	b	
20	0.243	0			a	
45	0.244	0.043	2	0.0215	b	
32	0.245	0			a	
33	0.2555	0.076	2	0.038	b	
14	0.260	0.048	2	0.024	b	
27	0.264	0.031	1.732	0.01789838	b	
91	0.267	0			a	
79	0.270	0.041	2	0.0205	b	
30	0.273	0.024	2	0.012	b	
4	0.274	0			a	
24	0.274	0			a	
110	0.278	0.004	2	0.002	b	
37	0.2811	0			a	
105	0.287	0.32	1.96	0.1633	c	
108	0.287	0			a	

43	0.288	0.021	2	0.0105	b	
75	0.289	0.03	2	0.015	b	
102	0.295	0			a	
99	0.296	0.087	2	0.0435	b	
40	0.311	0			a	
11	0.331	0.066	2	0.033	b	
68	**0.4246**	**0.15**	**3**	**0.05**	**b**	**Y**

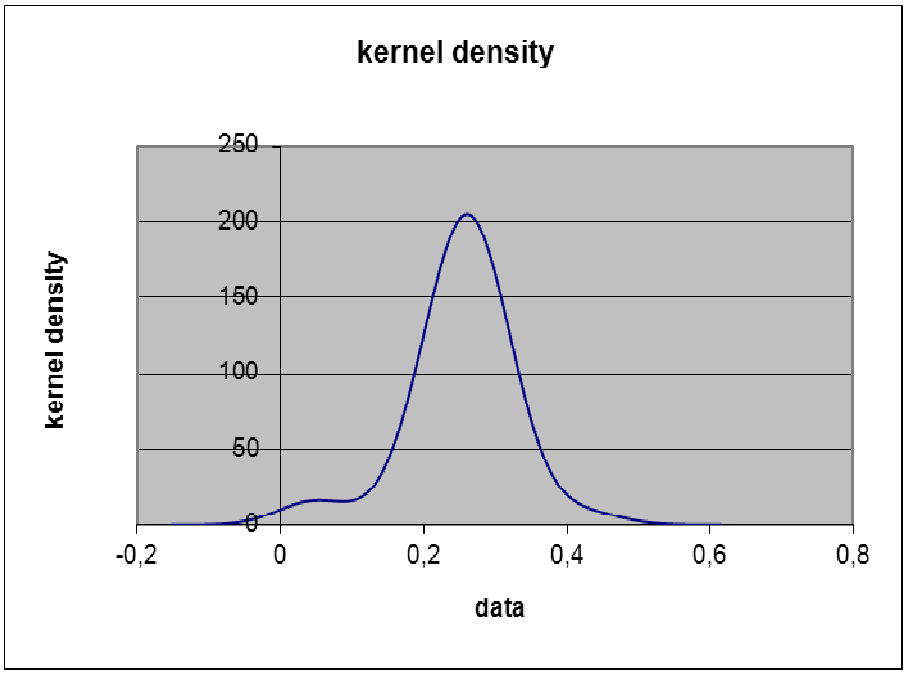


Figure E.2b — Kernel density plot for Atrazine data

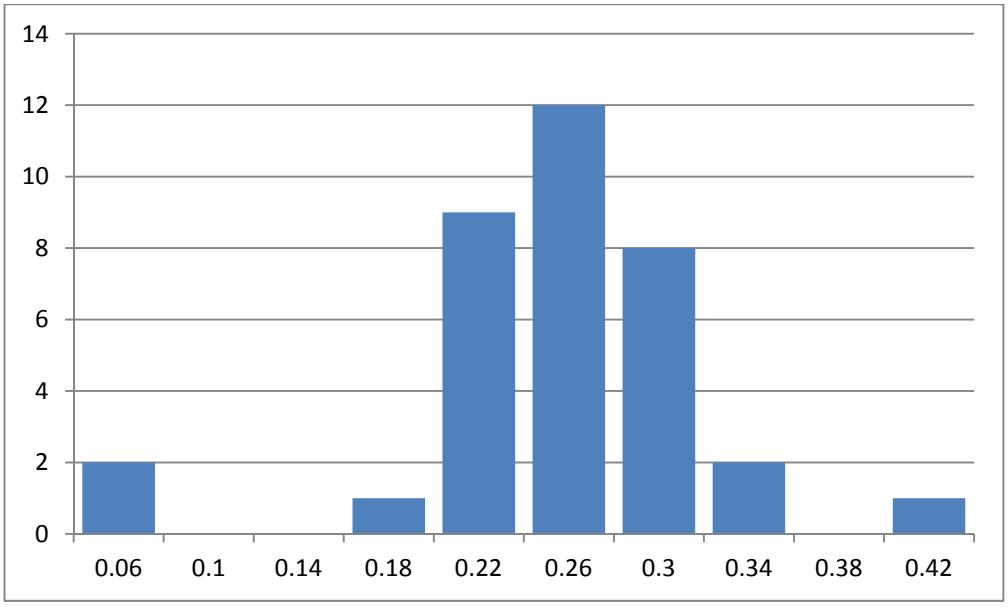


Figure E.2c — Histogram of participant results

For this study the assigned value was determined by gravimetric uncertainty, with components added from the homogeneity study and short term stability study.

The various components are u_{char} = 0,0018

u_{homo} =0,0040

u_{sts} = 0,0015

The uncertainty of the assigned value then is

$$u_{av} = \sqrt{u_{char}^2 + u_{homo}^2 + u_{sts}^2}$$

$u_{av} = \sqrt{((0{,}0018)^2 + (0{,}0040)^2 + (0{,}0015)^2)}$

u_{av} = 0,0046

Note for this example the uncertainty of the robust mean is calculated as follows:

$$u_X = 1{,}25\sqrt{s^{*2}/p}$$

$u_x = 1{,}25(\sqrt{(0{,}0082^2/36)}) = 0{,}0082$

It can be assumed that the uncertainty of the robust mean includes the effects of sample inhomogeneity and any effects of short term instability, including transportation.

E.3 Total Mercury in mineral feed (courtesy of IMEP®)

This example is from Study IMEP 111 from the Joint Research Center, Institute for Reference Materials and Measurements LINK IMEP-111 Report

Table E.3a — Proficiency test results from 24 participants in study IMEP 111

Lab code	Value	±	k	ulab	Flag	Method
L04	0,013	0,003	2	0,002	b	AMA
L05	0,013	0,007	2	0,004	a	AMA
L23	0,0135	0,00108	1,732	0,00062	b	AMA
L02	0,014	0,004	2	0,002	b	AMA
L15	0,014	0,0005	2	0,0003	b	AMA
L17	<0,015					CV-ICP-AES
L06	0,016	0,003	2	0,002	b	AMA
L09	0,017	0,008	2	0,004	a	AMA
L26	0,019	0,003	2	0,002	b	AAS

L12	0,0239	0,0036	2	0,0018	b	AMA
L13	<0,034					TDA-AAS
L03	0,037	0,013	2	0,007	a	CV-AAS
L29	0,039	0,007	2	0,004	a	CV-AAS
L07	0,04	0,008	2	0,004	a	ICP-MS
L21	0,04	0,03	2	0,02	c	HG-AAS
L25	0,040	0,010	2	0,005	a	CV-AAS
L16	0,0424	0,008	2	0,004	a	CV-AAS
L08	0,044	0,007	2	0,004	a	CV-AAS
L10	0,045	0,007	2	0,004	a	ICP-MS
L24	0,045	0,005	2	0,003	a	HG-AAS
L18	0,046	0,007	2	0,004	a	CV-AAS
L28	0,049	0,0072	2	0,0036	a	CV-AAS
L01	0,053	0,007	2	0,004	a	CV-AAS
L14	<0,1					ICP-MS

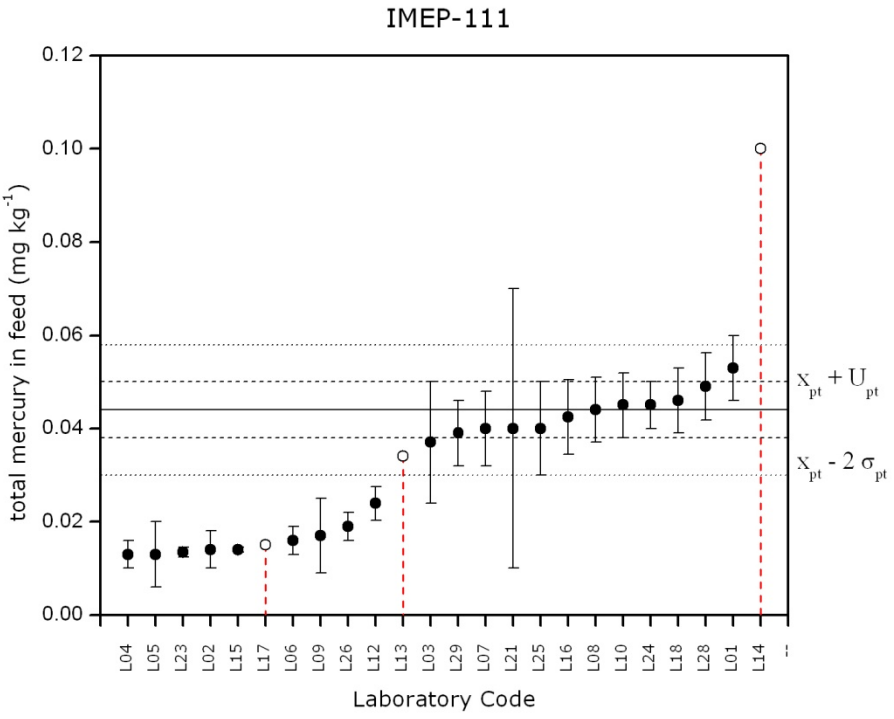


Figure E.3a — Participant results and uncertainties for results in IMEP 111

Table E.3b — Performance Statistics by various methods

Lab code	D%	P_A	z	z'	zeta	En
L04	-70%	-1,48	-4,43	-4,07	-9,24	-4,62
L05	-70%	-1,48	-4,43	-4,07	-6,72	-3,36
L23	-69%	-1,45	-4,36	-4,00	-9,95	-5,00
L02	-68%	-1,43	-4,29	-3,94	-8,32	-4,16
L15	-68%	-1,43	-4,29	-3,94	-9,97	-4,98
L17						
L06	-64%	-1,33	-4,00	-3,68	-8,35	-4,17
L09	-61%	-1,29	-3,86	-3,55	-5,40	-2,70
L26	-57%	-1,19	-3,57	-3,28	-7,45	-3,73
L12	-46%	-0,96	-2,87	-2,64	-5,75	-2,87
L13						
L03	-16%	-0,33	-1,00	-0,92	-0,98	-0,49
L29	-11%	-0,24	-0,71	-0,66	-1,08	-0,54
L07	-9%	-0,19	-0,57	-0,53	-0,80	-0,40
L21	-9%	-0,19	-0,57	-0,53	-0,26	-0,13
L25	-9%	-0,19	-0,57	-0,53	-0,69	-0,34
L16	-4%	-0,08	-0,23	-0,21	-0,32	-0,16
L08	0%	0,00	0,00	0,00	0,00	0,00
L10	2%	0,05	0,14	0,13	0,22	0,11
L24	2%	0,05	0,14	0,13	0,26	0,13
L18	5%	0,10	0,29	0,26	0,43	0,22
L28	11%	0,24	0,71	0,66	1,07	0,53
L01	20%	0,43	1,29	1,18	1,95	0,98
L14						

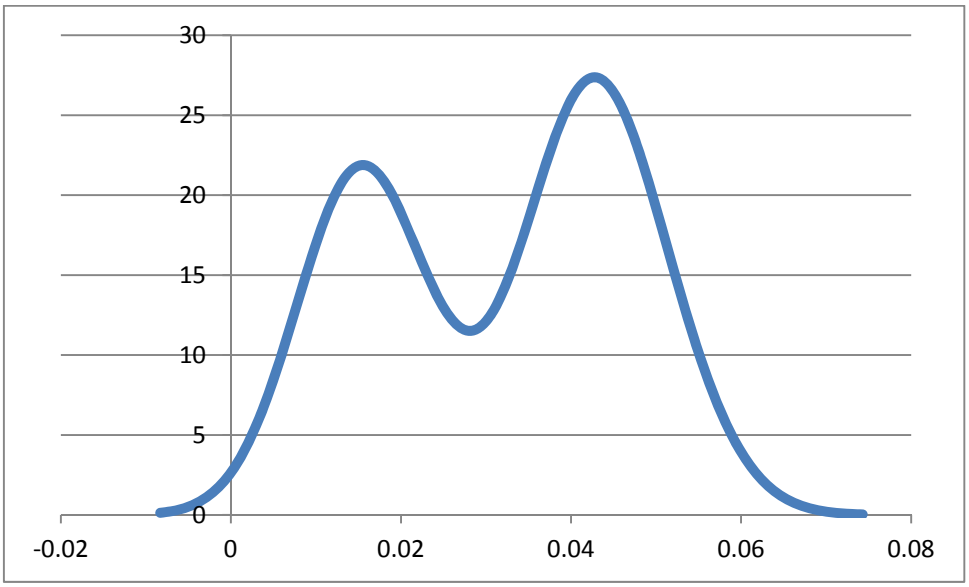


Figure E.3b — Kernel density plot for participant results in IMEP 111

E.4 Reference value from a single laboratory: Los Angeles value of aggregates

Table E.4 gives an example of data that might be obtained in such a series of tests, and shows how the standard uncertainty u_D of the difference is calculated. Note that uncertainty of the assigned value for the reference material includes the uncertainty due to inhomogeneity, transportation, and long term stability.

$$x_{pt} = 21{,}62 + 1{,}73 = 23{,}35\ LA\ units$$

And,

$$u(x_{pt}) = \sqrt{0{,}26^2 + 0{,}24^2} = 0{,}35 \text{ LA units}$$

where

0,26 is the standard uncertainty of the assigned value of the CRM, and 0,24 is the standard uncertainty of $\bar{d}$.

Table E.4 — Calculation of the average difference between a CRM and a proficiency test item, and of the standard uncertainty of this difference

Sample	RM		CRM		Difference in average values
	Test 1 LA units	Test 2 LA units	Test 1 LA units	Test 2 LA units	RM – CRM LA units
1	20,5	20,5	19,0	18,0	2,00
2	21,1	20,7	19,8	19,9	1,05
3	21,5	21,5	21,0	21,0	0,50
4	22,3	21,7	21,0	20,8	1,10
5	22,7	22,3	20,5	21,0	1,75
6	23,6	22,4	20,3	20,3	2,70
7	20,9	21,2	21,5	21,8	–0,60
8	21,4	21,5	21,9	21,7	–0,35
9	23,5	23,5	21,0	21,0	2,50
10	22,3	22,9	22,0	21,3	0,95
11	23,5	24,1	20,8	20,6	3,10
12	22,5	23,5	21,0	22,0	1,50
13	22,5	23,5	21,0	21,0	2,00
14	23,4	22,7	22,0	22,0	1,05

15	24,0	24,2	22,1	21,5	2,30
16	24,5	24,4	22,3	22,5	2,05
17	24,8	24,7	22,0	21,9	2,80
18	24,7	25,1	21,9	21,9	3,00
19	24,9	24,4	22,4	22,6	2,15
20	27,2	27,0	24,5	23,7	3,00
Average difference, $\bar{d}$					1,73
Standard deviation					1,07
Standard uncertainty of $\bar{d}$ (standard deviation / $\sqrt{20}$)					0,24
NOTE The data are measurements of the mechanical strength of aggregate, obtained from the Los Angeles (LA) test.					

E.5 Results from a study using expert laboratories (section 7.6)

A study by the International Measurement Evaluation program (IMEP), study 112 was to test for Total Arsenic (As) in Wheat. In this study the assigned value was determined as the combined results from a group of 7 highly reputable laboratories, each testing 2 units of the proficiency testing item. The laboratories reported their measurement method, all replicate results, and their claimed expanded measurement uncertainty (k=2) of this for that level of As in wheat.

Table E.5 — Reference results from Expert Laboratories for Total Arsenic in Wheat

Expert	**Average**	U_X **(*k*=2)**
E1	0,188	0,024
E2	0,178	0,008
E3	0,195	0,037
E4	0,157	0,005
E5	0,175	0,003
E6	0,179	0,011
E7	0,166	0,009

The robust mean from Algorithm A is 0,177, which is the same as the arithmetic mean
The uncertainty of the robust mean is calculated according to equation (6) as

u_{char} = 1,25/7 * $\sqrt{\Sigma(U_i / 2)^2}$ = 0,1786 * $\sqrt{}$(0,000561) = 0,0042

This is very close to the value determined by the proficiency testing provider, calculated with a different formula (from ISO Guide 35):

$u_{char} = SD_{means}/\sqrt{7}$ = 0,0048 with SD_{means} = standard deviation of mean values from 7 laboratories

E.6 Determination of evaluation criteria by experience with previous rounds (section 8.3): toxaphene in drinking water

There are two proficiency testing providers for the pesticide Toxaphene (a pesticide) in drinking water. Over a period of 5 years there have been 20 rounds of PT where there were 20 or more participants, covering regulated Toxaphene levels from 3 to 20 μg/L. Table E.6 shows the results from the 20 studies, arranged from low to high assigned values. Figures E.6a and E.6b show the scatter plots for the robust standard

deviation (SD) and relative robust standard deviation (RSD%) for each study, compared with the assigned value (from formulation). The formulae for the simple least-squares linear regression line are shown for each figure. Least squares regression lines can be determined with generally available spread-sheet software.

It is apparent that the RSD is fairly constant at about 19% for all levels, and that the regression line for SD is reasonably reliable (R^2 = 0,83). A regulatory body may choose to require that the standard deviation for proficiency assessment be 19% of the assigned value (or perhaps 20%), or they may require calculation of the expected standard deviation, using the regression equation for SD.

Table E.6 — Proficiency testing rounds for Toxaphene in drinking water and p≥20 results

PT Provider Code	Assigned Value	Robust Mean	Standard Deviation	Mean Recovery	STDEV (% of AV)	p
P004	3.96	3.98	0.639	100.5%	16.1%	25
P001	4.56	5.18	0.638	113.6%	14.0%	23
P001	5.99	5.98	0.995	99.8%	16.6%	22
P004	6.08	5.8	1.48	95.4%	24.3%	20
P001	6.2	6.66	0.971	107.4%	15.7%	23
P001	6.72	7.13	1.43	106.1%	21.3%	22
P004	8.1	7.09	2.23	87.5%	27.5%	21
P001	8.73	8.15	1.8	93.4%	20.6%	22
P001	9.57	8.6	1.45	89.9%	15.2%	23
P001	12.1	12.4	1.44	102.5%	11.9%	23
P001	12.5	13.8	2.25	110.4%	18.0%	24
P004	13.1	12	2.41	91.6%	18.4%	20
P004	15.6	13.3	3.57	85.3%	22.9%	27
P004	15.9	13.6	2.44	85.5%	15.3%	28
P004	16.3	13.5	3.6	82.8%	22.1%	31
P004	16.3	14.2	3.09	87.1%	19.0%	40
P004	17	15.6	2.63	91.8%	15.5%	24
P004	17.4	16	2.85	92.0%	16.4%	23
P004	17.4	16	3.36	92.0%	19.3%	23
P004	19	16.4	3.2	86.3%	16.8%	27

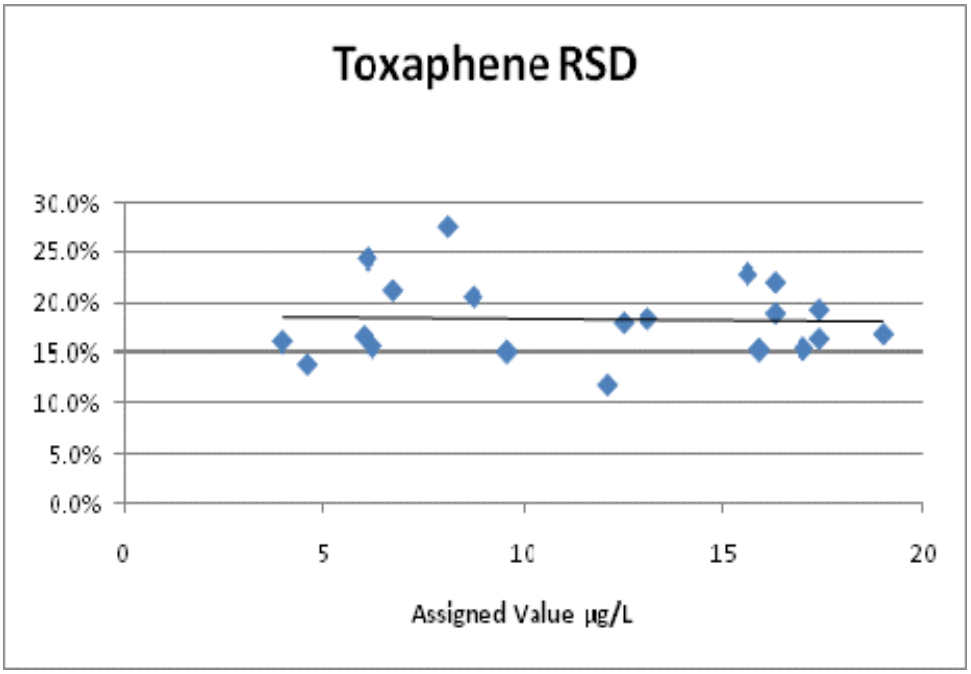


Figure E.6a — RSD participant results (%) vs Concentration (µg/L)

Regression equation: RSD = -0,0011(AV) + 0,1957
p(slope)=0,60 R^2 = 0,02

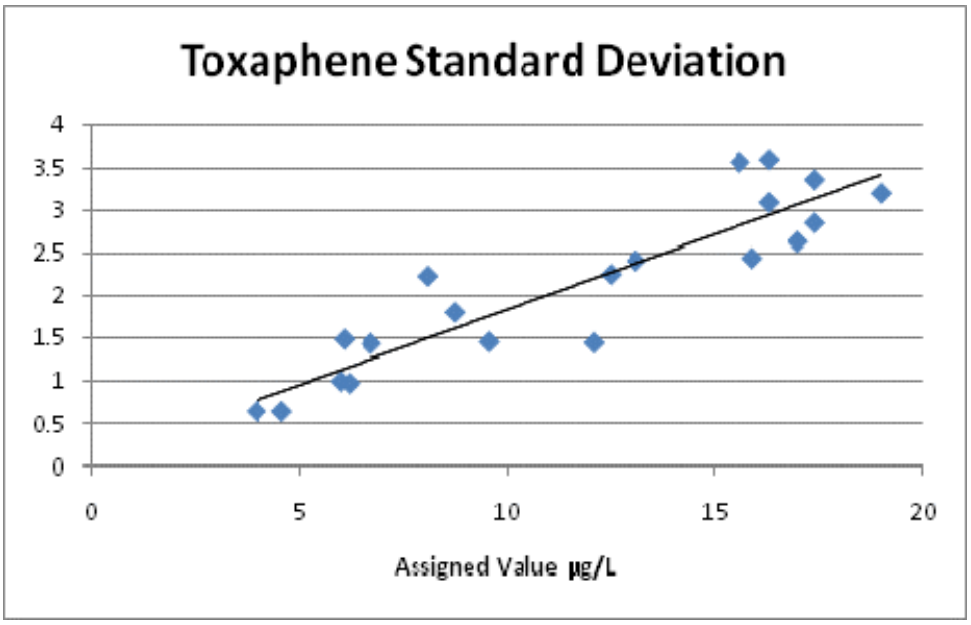


Figure E.6b — Participant Standard Deviation vs Concentration (µg/L)

Regression equation: SD=0,1972$^{x}\boldsymbol{x_{pt}}$ - 0,0003

p(slope)<0,001 R^2 = 0,74

Polynomial quadratic model, 1st and 2nd order coefficients not significant.

E.7 From a general model (section 8.4.2.1): Horwitz equation

One common general model for chemical applications was described by Horwitz [20]. This approach gives a general model for the reproducibility of analytical methods that may be used to derive the following expression for the reproducibility standard deviation:

$$\sigma_R = 0{,}02 \times c^{0{,}8495}$$

where

c is the concentration of the chemical species to be determined in mass fraction.

For example, a proficiency test of melamine in milk powder uses two proficiency test samples with reference levels A= 1,195 mg/kg and B= 2.565 mg/kg (.0001195% and .0002565%). This yields the following expected reproducibility standard deviations.

Sample A at 1,195 mg/kg: σ_R = 0,046 mg/kg or relative σ_R = 3.9%

Sample B at 2,565 mg/kg: σ_R = 0,089 mg/kg or relative σ_R = 3.5%

E.8 Determining performance from a precision experiment: Determination of the cement content of hardened concrete

The cement content of concrete is usually measured in terms of the mass in kilograms of cement per cubic metre of concrete (i.e. in kg/m^3). In practice, concrete is produced in grades of quality that have cement contents 25 kg/m^3 apart, and it is desirable that participants should be able to identify the grade correctly. For this reason, it is desirable that the chosen value of σ_{pt} should be no more than one-half of 25 kg/m^3. A precision experiment produced the following results, for a concrete with an average cement content of 260 kg/m^3: σ_R = 23,2 kg/m^3 and σ_r = 14,3 kg/m^3.

So

$$\sigma_L = \sqrt{23{,}2^2 - 14{,}3^2} = 18{,}3 \text{ kg/m}^3$$

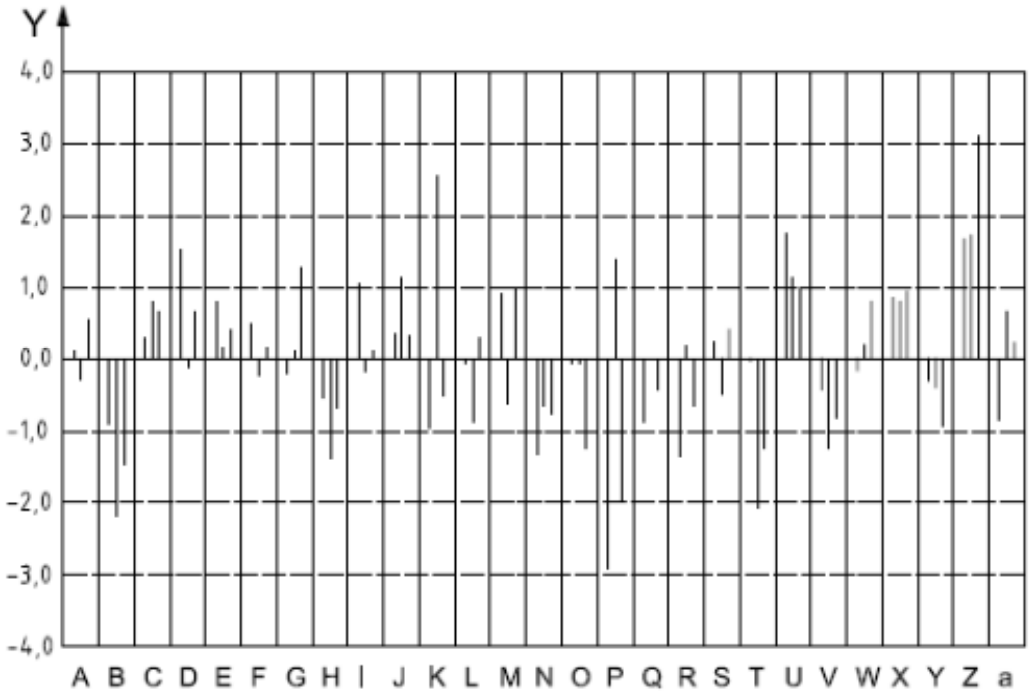
Taking n as 2, and substituting σ_L = 18,3 kg/m^3, σ_r = 14,3 kg/m^3 gives:

$$\hat{\sigma} = \sqrt{18{,}3^2 + \left(14{,}3^2/2\right)} = 20{,}9 \text{ kg/m}^3$$

assuming that m = 2 replicate measurements are to be made.

E.9 Bar-plots of standardized biases: Antibody concentrations

The z-scores from a study with three related measurands (antibodies) from are shown plotted as a bar-chart. From this graph, laboratories B and Z (for example) can see that they should look for a cause of bias that affects all three levels by approximately the same amount, whereas laboratories K and P (for example) can see that in their case the sign of the z-score depends on the type of antibody.



Key

Y z-score

NOTE "A" to "a" indicates the laboratory number.

Figure E.9 — Bar-chart of *z*-scores (4,0 to –4,0) for one round of a proficiency test in which the participants determined the concentrations of three allergen specific IgE antibodies

E.10 Youden Plot

E.10.1 Confidence ellipse

Call the two proficiency test items A and B, and denote the results obtained on A by:

$x_{A,1}, x_{A,2}, \ldots, x_{A,p}$

and those obtained on B by:

$x_{B,1}, x_{B,2}, \ldots, x_{B,p}$

where

p is the number of participants.

Calculate the averages and standard deviations of the two sets of data:

$\bar{x}_{A,.}\ \bar{x}_{B,.}\quad s_A\ s_B$

and the rank correlation coefficient $\hat{\rho}$. Calculate the *z*-scores for the two materials as follows (note, this may be different than z scores used for performance evaluation):

$$z_{A,i} = \left(x_{A,i} - \bar{x}_{A,.}\right) \Big/ s_A \quad \text{where } i = 1, 2, \ldots, p \tag{E.1}$$

$$z_{B,i} = \left(x_{B,i} - \bar{x}_{B,.}\right) \Big/ s_B \quad \text{where } i = 1, 2, ..., p \tag{E.2}$$

and calculate the combined scores for the two proficiency test items:

$$z_{A,B,i} = \sqrt{z_{A,i}^2 \; - \; 2\hat{\rho}\, z_{A,i} z_{B,i} \; + \; z_{B,i}^2} \tag{E.3}$$

Define standardized variables as:

$$z_A = \left(x_A - \bar{x}_{A,.}\right) \Big/ s_A \tag{E.4}$$

$$z_B = \left(x_B - \bar{x}_{B,.}\right) \Big/ s_B \tag{E.5}$$

In terms of the standardized variables, the confidence ellipse may be written in terms of Hotelling's T^2:

$$z_A^2 - 2\hat{\rho} z_A z_B + z_B^2 = \left(1 - \hat{\rho}^2\right) T^2 \tag{E.6}$$

where

$$T^2 = 2\left\{\left(p-1\right)\Big/\left(p-2\right)\right\} F_{(1-\alpha)}\left(2, p-1\right) \tag{E.7}$$

Here, $F_{(1\,-\,\alpha)}(2,\ p\ -\ 1)$ is the tabulated $(1\ -\ \alpha)$-fractile of the F-distribution with 2 and $(p\ -\ 1)$ degrees of freedom. The ellipse may be drawn on a graph that has the z-scores z_A and z_B as the axes by plotting a series of points for $-T \; u \; z_A \; u \; T$ with:

$$z_B = \hat{\rho}\, z_A \pm \sqrt{\left(1 \; - \; \hat{\rho}^2\right)\left(T^2 \; - \; z_A^2\right)} \tag{E.8}$$

NOTE 1 To plot the confidence ellipse on a graph with axes that show the original units of measurement, transform the above series of points back to the original units using:

$$x_A = \bar{x}_{A,.} + s_A \times z_A$$

$$x_B = \bar{x}_{B,.} + s_B \times z_B$$

To plot the confidence ellipse on a graph with axes that show differences D_A and D_B, transform the above series of points using:

$$D_A = s_A \times z_A$$

$$D_B = s_B \times z_B$$

To plot the confidence ellipse on a graph with axes that show percentage differences $D_{A\%}$ and $D_{B\%}$, transform the above series of points using:

$$D_{A\%} = 100 s_A \times z_A / x_A$$

$$D_{B\%} = 100 s_B \times z_B / x_B$$

The combined *z*-scores may be used as an aid to interpreting the Youden Plot. The highest combined *z*-scores correspond to the highest significance levels 100α % in the calculation of the confidence ellipse, so the combined *z*-scores may be used to identify the most extreme points on the Youden Plot. On occasion, it may be necessary to exclude one or more outlying points and re-calculate the ellipse: the combined *z*-scores may then be used as an aid to identifying the points to exclude.

NOTE 2 There is a need for a robust method of calculating the ellipse, but the details of such a method have not yet been worked out. The cut-off value may be calculated by noting that $(z_{A,B,i})^2 / (1 - \hat{\rho}^2)$ has approximately the chi-squared distribution with 2 degrees of freedom, but the correction factor may have to be derived by simulation.

E.10.2 Example; antibody concentrations

Table E.10 shows data obtained by testing two similar samples for antibody concentrations, and the calculations required to derive the confidence ellipse. With p = 29 laboratories, and using a significance level of 100α % = 5 %, $F_{(1-\alpha)}(2, p-1)$ = 3,34. Hence T = 2,632 and, in terms of the standardized variables, the 95 % confidence ellipse may be written:

$$z_A^2 - 1{,}412 z_A z_B + z_B^2 = 3{,}48 \qquad \text{(E.9)}$$

The ellipse is shown, together with the points representing the *z*-scores, in Figure E.10, together with the ellipses for probability levels of 100α % = 1 % and 0,1 %. The combined scores are shown in Table E.10.

Inspection of Figure E.10 reveals two laboratories (numbers 5 and 23) in the top right-hand quadrant. They have combined *z*-scores of 1,641 and 2,099. Laboratory 26 has a high *z*-score on item B and a combined *z*-score of 2,059. After laboratories 5, 23 and 26, the laboratory that gives the next highest combined *z* score is number 8 (combined score of 1,501).

The points for laboratories 23 and 26 fall between the ellipses for the 5 % and 1 % probability levels, so it would be appropriate to treat their results as giving rise to "warning" signals, and to check where their results fall in the next round of the scheme.

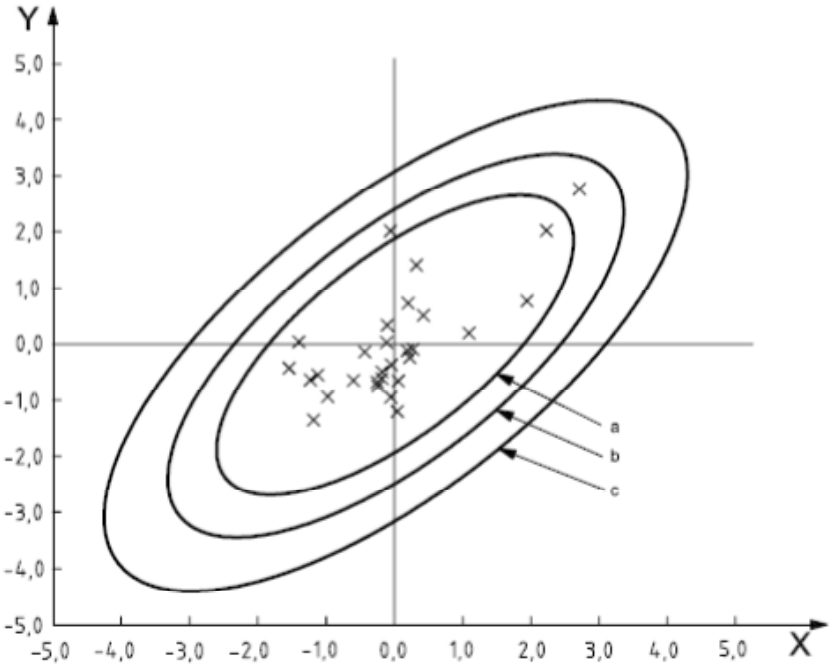
Table E.10 — Data and calculations on concentrations of antibodies for two similar allergens

Laboratory	Data		*z*-score		Combined score
	Allergen A	Allergen B	Allergen A	Allergen B	
i	$x_{A,i}$	$x_{B,i}$	$z_{A,i}$	$z_{B,i}$	$z_{A,B,i}$
1	12,95	9,15	0,427	0,515	0,370
2	6,47	6,42	–1,540	–0,428	1,275
3	11,40	6,60	–0,043	–0,366	0,336
4	8,32	4,93	–0,978	–0,942	0,737
5	18,88	13,52	2,228	2,023	1,641
6	15,14	8,22	1,092	0,194	0,965
7	10,12	7,26	–0,432	–0,138	0,349
8	17,94	9,89	1,942	0,770	1,501
9	11,68	4,17	0,042	–1,204	1,234
10	12,44	7,39	0,272	–0,093	0,344
11	6,93	7,78	–1,400	0,042	1,430
12	9,57	5,80	–0,599	–0,642	0,477
13	11,73	5,77	0,057	–0,652	0,693
14	12,29	6,97	0,227	–0,238	0,429
15	10,95	6,23	–0,180	–0,493	0,388
16	10,95	5,90	–0,180	–0,607	0,497
17	11,17	7,74	–0,113	0,028	0,134
18	11,20	8,63	–0,104	0,335	0,415
19	7,64	3,74	–1,185	–1,353	0,986
20	12,17	7,33	0,190	–0,114	0,282

21	10,71	5,70	–0,253	–0,676	0,529
22	7,84	6,07	–1,124	–0,549	0,833
23	20,47	15,66	2,710	2,762	2,099
24	12,60	11,76	0,321	1,415	1,210
25	11,37	4,91	–0,052	–0,949	0,913
26	11,36	13,51	–0,055	2,019	2,059
27	10,75	5,48	–0,241	–0,752	0,607
28	12,21	9,77	0,203	0,729	0,603
29	7,49	5,82	–1,230	–0,635	0,902
Average	11,54	7,66	0,00	0,00	
Standard deviation	3,29	2,90	1,00	1,00	
Correlation coefficient	0,706		0,706		

NOTE 1 The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

NOTE 2 The *z*-scores in this table have been calculated using non-rounded values of the averages and standard deviations, not using the rounded values shown at the bottom of the table.



Key

X *z*-score for allergen A
Y *z*-score for allergen B
a 5 % level.
b 1 % level.
c 0,1 % level.

Figure E.10 — Youden Plot of *z*-scores from Table E.10

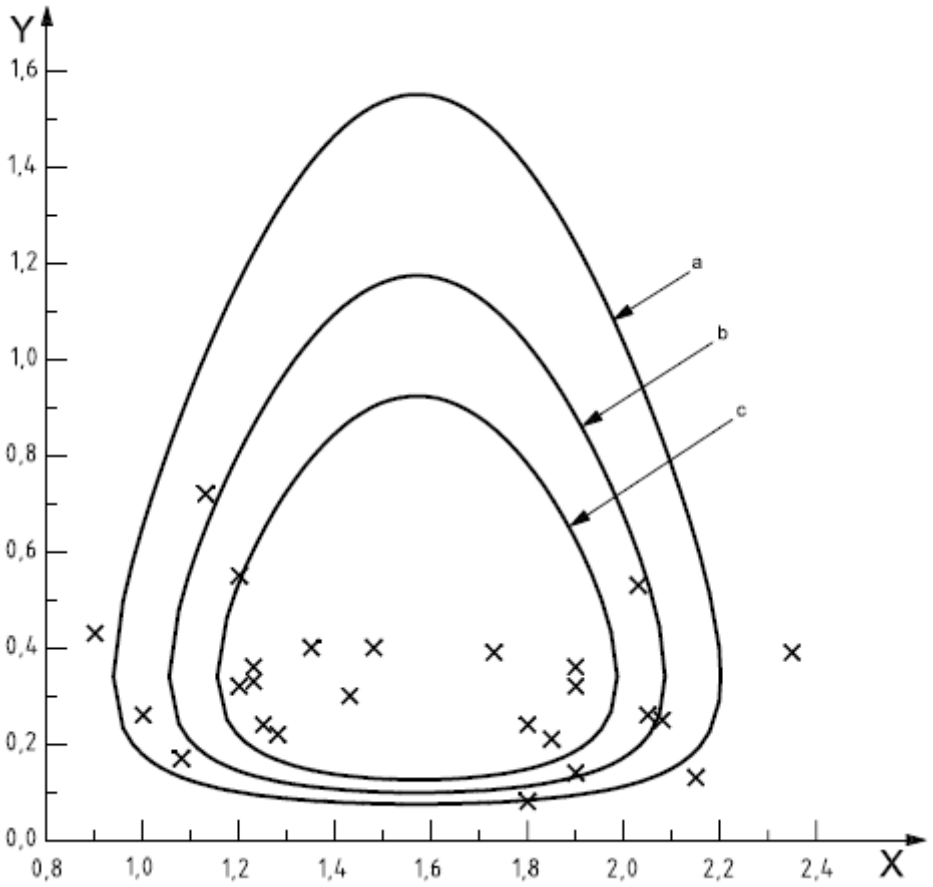
E.11 Plot of repeatability standard deviations: Antibody concentrations

Table E.11 shows the results of determining concentrations of a certain antibody in serum samples. Each laboratory made four replicate determinations, under repeatability conditions. The formulae given above are used to obtain the plot shown as Figure E.11. The plot shows that several of the laboratories receive action or warning signals.

Table E.11 — Concentrations of certain antibodies in serum samples (four replicate determinations on one sample in each laboratory)

Laboratory	Average kU/l	Standard deviation kU/l
1	2,15	0,13
2	1,85	0,21
3	1,80	0,08
4	1,80	0,24
5	1,90	0,36
6	1,90	0,32
7	1,90	0,14
8	2,05	0,26
9	2,35	0,39
10	2,03	0,53
11	2,08	0,25
12	1,25	0,24
13	1,13	0,72
14	1,00	0,26
15	1,08	0,17
16	1,20	0,32
17	1,35	0,4
18	1,23	0,36
19	1,23	0,33
20	0,90	0,43
21	1,48	0,40
22	1,20	0,55
23	1,73	0,39
24	1,43	0,30
25	1,28	0,22
Robust average	1,57	
Robust standard deviation		0,34

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.



Key
X average
Y standard deviation

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

a 0,1 % level.

b 1 % level.

c 5 % level.

Figure E.11 — Plot of standard deviations against averages for 25 laboratories (data from Table E.11)

E.12 Split samples: Antibody concentrations

The concentration of certain antibodies in 21 serum samples were measured with radioimmunoassay methods in two laboratories denoted X and Y. In each laboratory all measurements were performed in duplicate in the same run. The concentrations obtained (in U/l) are presented in Table E.12a. As the measurement range is large, relative differences are relevant, so the data are transformed by taking logarithms to base e before the calculations are performed. The transformed data are shown in Table E.12a and graphs showing the statistics from Table E.12a are shown in Figures E.12a – E.12b

From the graphs of ranges of replicate determinations, it appears that the variation between replicates for laboratory X is higher than for laboratory Y. Pooled values of these statistics are shown in Table E.12b and could be compared using an *F*-test if it was of interest. Looking at the third graph, it can be seen that there is no obvious pattern or trend in the points. However, whereas the ranges of replicate determinations in Figures E.12a and E.12b are nearly all less than 0,2, many of the differences between laboratories in Figure E.12c are much larger than this. This aspect requires investigation because it implies that the

difference between the laboratories depends on the sample. The average difference between the laboratories may be calculated and is shown in Table E.12a. It may be used to give an indication of the importance of the difference between the laboratories, but it may not be used to predict the difference between the laboratories that might be obtained when analysing some subsequent sample. Thus with the transformed data, on average $\ln(Y) - \ln(X) = 0{,}443$, so $Y/X = 1{,}6$, indicating that laboratory Y obtains results, on average, higher than laboratory X by a factor of 1,6. However, with some samples the difference is much larger, and on others laboratory X obtains the higher results.

Table E.12a — Concentrations of certain antibodies in 21 serum samples

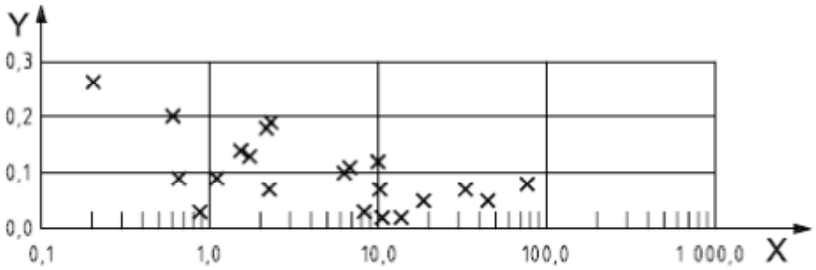
Sample	Laboratory X		Laboratory Y		Laboratory X	Laboratory Y	Laboratories X and Y
i	Replicate 1 U/l	Replicate 2 U/l	Replicate 1 U/l	Replicate 2 U/l	Average U/l	Average U/l	Average U/l
1	19,106	18,174	11,473	11,705	18,640	11,589	15,115
2	6,424	7,171	5,812	5,812	6,798	5,812	6,305
3	6,619	5,989	11,705	11,473	6,304	11,589	8,947
4	0,543	0,664	0,861	0,905	0,604	0,883	0,743
5	43,816	46,063	49,899	55,147	44,940	52,523	48,731
6	2,096	2,535	24,047	26,843	2,316	25,445	13,880
7	10,591	9,875	9,116	8,671	10,233	8,894	9,563
8	13,874	13,599	12,554	12,807	13,737	12,681	13,209
9	1,974	2,363	1,094	1,020	2,169	1,057	1,613
10	9,393	10,591	13,736	14,585	9,992	14,161	12,076
11	1,840	1,616	2,484	2,460	1,728	2,472	2,100
12	31,817	34,124	48,424	55,147	32,971	51,786	42,378
13	1,150	1,051	2,014	2,270	1,101	2,142	1,621
14	0,625	0,684	1,051	1,174	0,655	1,113	0,884
15	73,700	79,838	119,104	127,740	76,769	123,422	100,096
16	2,181	2,340	2,560	3,065	2,261	2,813	2,537
17	8,415	8,166	5,755	5,585	8,291	5,670	6,980
18	1,419	1,632	8,846	8,846	1,526	8,846	5,186
19	0,861	0,887	2,612	3,065	0,874	2,839	1,856
20	10,697	10,486	15,029	14,880	10,592	14,955	12,773
21	0,230	0,177	0,795	0,795	0,204	0,795	0,499

NOTE The data are numbers of units (U) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

Table E.12b — ln(concentrations) and statistics for the data in Table E.12a

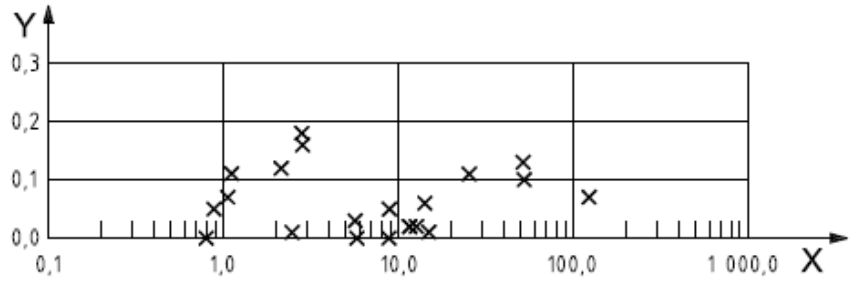
Sample	Laboratory X		Laboratory Y		Laboratory X	Laboratory Y	Difference Y – X
i	Replicate 1 ln U/l	Replicate 2 ln U/l	Replicate 1 ln U/l	Replicate 2 ln U/l	Range ln U/l	Range ln U/l	ln U/l
1	2,95	2,90	2,44	2,46	0,05	0,02	–0,475

2	1,86	1,97	1,76	1,76	0,11	0,00	–0,155
3	1,89	1,79	2,46	2,44	0,10	0,02	0,610
4	–0,61	–0,41	–0,15	–0,10	0,20	0,05	0,385
5	3,78	3,83	3,91	4,01	0,05	0,10	0,155
6	0,74	0,93	3,18	3,29	0,19	0,11	2,400
7	2,36	2,29	2,21	2,16	0,07	0,05	–0,140
8	2,63	2,61	2,53	2,55	0,02	0,02	–0,080
9	0,68	0,86	0,09	0,02	0,18	0,07	–0,715
10	2,24	2,36	2,62	2,68	0,12	0,06	0,350
11	0,61	0,48	0,91	0,90	0,13	0,01	0,360
12	3,46	3,53	3,88	4,01	0,07	0,13	0,450
13	0,14	0,05	0,70	0,82	0,09	0,12	0,665
14	–0,47	–0,38	0,05	0,16	0,09	0,11	0,530
15	4,30	4,38	4,78	4,85	0,08	0,07	0,475
16	0,78	0,85	0,94	1,12	0,07	0,18	0,215
17	2,13	2,10	1,75	1,72	0,03	0,03	–0,380
18	0,35	0,49	2,18	2,18	0,14	0,00	1,760
19	–0,15	–0,12	0,96	1,12	0,03	0,16	1,175
20	2,37	2,35	2,71	2,70	0,02	0,01	0,345
21	–1,47	–1,73	–0,23	–0,23	0,26	0,00	1,371
Pooled range Average difference between the laboratories					0,119	0,083	 0,443
NOTE The data are numbers of units (U) per litre (l) of sample, where a unit is defined by the concentration of an international reference material. The pooled range is calculated according to Algorithm S in Annex C.							



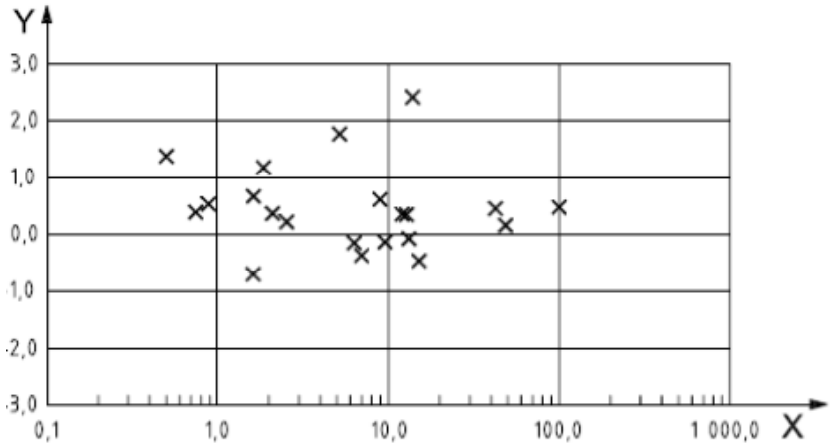
Key
X average concentration for laboratory X, %
Y range of ln(concentrations) for laboratory X

Figure E.12a — Ranges of replicate determinations for laboratory X (with the ranges calculated from the in concentrations)



Key
X average concentration for laboratory Y, %
Y range of ln(concentrations) for laboratory Y

Figure E.12b — Ranges of replicate determinations for laboratory Y (with the ranges calculated from the in concentrations)



Key
X average concentration for laboratories X and Y, %
Y difference in ln(concentrations) laboratory Y – laboratory X

Figure E.12c — Differences between laboratory averages Y – X (with the differences calculated from the in concentrations)

E.13 Graphical methods for tracking performance over time

It can be useful for a laboratory to track their own performance over time, or to have this prepared by the proficiency testing provider. A simple and conventional tool is a quality control chart, or Shewhart plot. This requires having a standardized performance score, such as z score or P_A score and participation over some event. This example is from a medical EQA programme, for serum potassium.

This PT provider uses a fixed interval for acceptance of 5%, although with rounding to next reportable value (0,1 mmol/L), and no smaller than +/- 0,2 mmol/L. The provider uses P_A scores rather than z scores.

Table E.13 — P_A scores for 5 Rounds of EQA, each with 3 samples for Serum Potassium

Event Code	Sample	Result	Assigned Value	P_A Score	Avg P_A
101	A	6.4	6.2	75	42
101	B	4.2	4.1	50	

101	C	4.1	4.1	0	
102	A	6	5.9	25	8
102	B	4.3	4.4	-33	
102	C	5.5	5.4	33	
103	A	4.1	4.2	-33	-28
103	B	3.6	3.7	-50	
103	C	4.2	4.2	0	
104	A	5.7	5.8	-25	11
104	B	3.9	4	-50	
104	C	6.3	5.9	110	
105	A	3.6	3.7	-50	-19
105	B	4.5	4.6	-33	
105	C	5.3	5.2	25	

The results can easily be plotted for review – 2 types of plots are recommended:

Quality control chart of the standardized score for each round, showing multiple samples in the same round. This will highlight performance over time, including any trends. This is Figure E.13a.

Scatter plot of standardized scores against assigned value, to see if performance is related to concentration level, again, to show any trends related to level of the measurand. This is Figure E.13b.

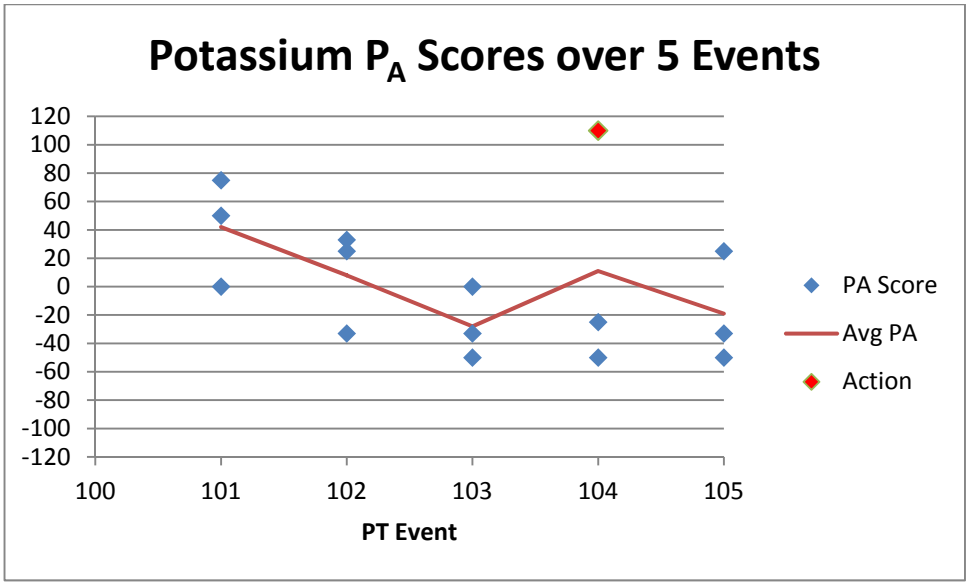


Figure E.13a — Scores for each event

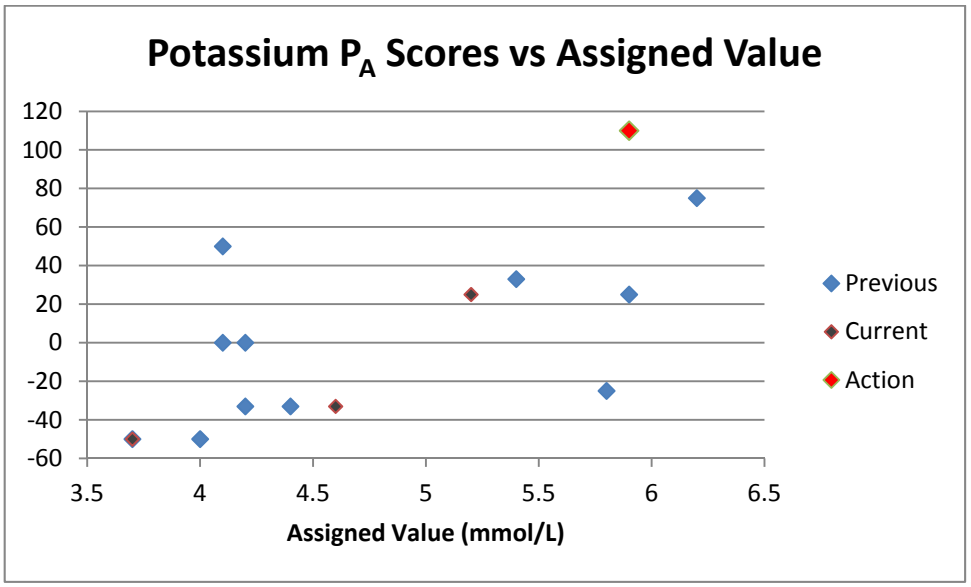


Figure E.13b — Scores for different levels of the measurand

E.14 Qualitative Data Analysis from Section 11.4; Example of an ordinal quantity: skin reaction to a cosmetic

Considerations for specific situations, such as determining the assigned value, summary statistics and graphical display are discussed in the following example. In this example a proficiency testing scheme involves the analysis of the response to a skin care product, when applied to a standard animal subject (species and type). Any inflammatory response is graded according to the following scale:

a) No response
b) moderate redness
c) significant irritation or swelling
d) severe reaction, including suppuration or bleeding

Two products are distributed, labelled products A and B, and there are 50 participants for each product. The responses are listed below and, for brevity, the mode and median are listed for each sample.

Table E.14 — Results for Two samples, skin irritation

Response	Product A	Product B
1	20 (40%) *	8 (16%)
2	18 (36%) **	12 (24%)
3	10 (20%)	20 (40%) * **
4	2 (4%)	10 (20%)

* indicates modal response

** indicates median response

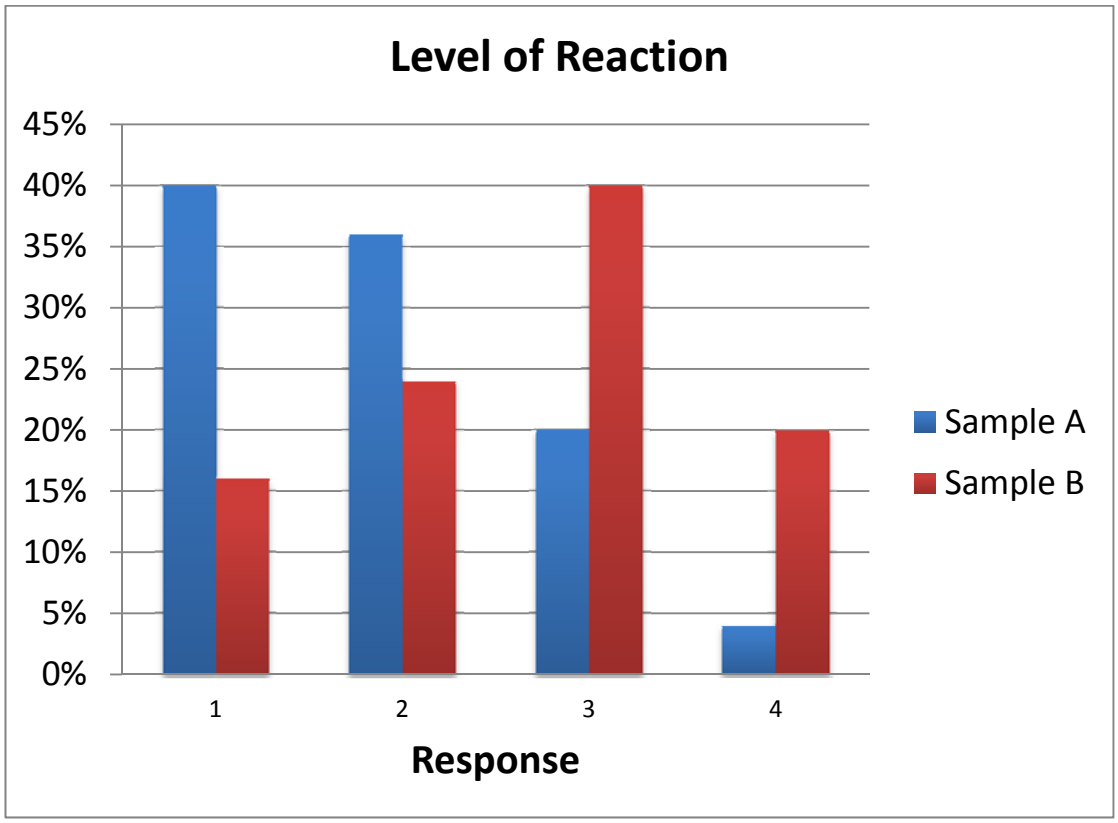


Figure E.14 — Bar chart of percentage responses to 2 skin irritation 0roficiency testing samples

Note that the median or mode may be used as summary statistics for these samples, and they suggest that the level of response to sample B was more severe than the response to sample A. The proficiency test provider may determine that "action signals" would occur for any result that is more than one ordinal unit away from the median, in which case for sample A, action signals occur for the 2 results of "4" and for sample B, action signals occur for the 8 results of "1".

E.15 Homogeneity and Stability test – Arsenic in animal feed

A material is prepared for use in an international proficiency test, and then subsequent use as a reference material. 1000 vials are manufactured and 10 samples are selected using a stratified random selection of samples from different portions of the manufacture process. 2 test portions are extracted from each bottle and tested in a random order, under repeatability conditions. The data are given in Table E.15 below. The procedure in Annex B.3 are followed, resulting in the summary statistics listed. The fitness-for purpose σ_{pt} for As in Feed is 15%, so the estimate of sample variability is checked against 0,3 times the σ_{pt}.

Table E.15a Homogeneity data for proficiency test of arsenic in animal feed

Bottle ID	Replicate 1	Replicate 2
3	0.185	0.194
111	0.187	0.189
201	0.182	0.186
330	0.188	0.196
405	0.191	0.181

481	0.188	0.18
599	0.187	0.196
704	0.177	0.186
766	0.179	0.187
858	0.188	0.196

Overall average: 0,18715
SD of averages: 0,00398
s_w: 0,00556
s_s: 0,00060
σ_{pt} =0,18715*0,15 = 0,02807
Check value: 0,3*σ_{pt} = 0,00842
s_s < Check value

Stability check: 2 samples are randomly selected and stored at an elevated temperature (60C) for the duration of the study (6 weeks). The samples were tested in duplicate, and the four results are checked against the homogeneity values.

Table E.15b Stability data for proficiency test of arsenic in animal feed

Stability sample	Replicate 1	Replicate 2
164	0.191	0.198
732	0.19	0.196

Overall average = 0,19375
Difference from Homogeneity mean: 0,19375 - 0,18715 = 0,00660
Check value 0,3*σ_{pt} = 0,00842 Sufficient stability.

Bibliography

[1] ISO 5725-2, *Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method*

[2] ISO 5725-3, *Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method*

[3] ISO 5725-4, *Accuracy (trueness and precision) of measurement methods and results — Part 4: Basic methods for the determination of the trueness of a standard measurement method*

[4] ISO 5725-5, *Accuracy (trueness and precision) of measurement methods and results — Part 5: Alternative methods for the determination of the precision of a standard measurement method*

[5] ISO 5725-6, *Accuracy (trueness and precision) of measurement methods and results — Part 6: Use in practice of accuracy values*

[6] ISO 8258, *Shewhart control charts*

[7] ISO 11352, *Water quality – Estimation of measurement uncertainty based on validation and quality control data*

[8] ISO 11843-1, *Capability of detection — Part 1: Terms and definitions*

[9] ISO 11843-2, *Capability of detection — Part 2: Methodology in the linear calibration case*

[10] ISO 16269-4, *Statistical interpretation of data — Part 4: Detection and treatment of outliers*

[11] ISO/IEC 17011, *Conformity assessment — General requirements for accreditation bodies accrediting conformity assessment bodies.*

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

[13] ISO Guide 34, *General requirements for the competence of reference material producers*

[14] ISO Guide 35, *Reference materials — General and statistical principles for certification*

[15] Analytical Method Committee, Royal Society of Chemistry Accred Qual Assur 15:73–79, 2010

[16] Efron B, Tibshirani R, An Introduction to the Bootstrap. Chapman & Hall, 1993

[17] Davison A C, Hinkley D V, Bootstrap Methods and Their Application. Cambridge University Press, 1997

[18] Gower J C, A general coefficient of similarity and some of its properties, Biometrics Vol. 27, No. 4, pp. 857-871, 1971

[19] Helsel D R, Nondetects and data analysis: statistics for censored environmental data. Wiley Interscience, 2005

[20] Horwitz W, Evaluation of analytical methods used for regulations of food and drugs. *Anal. Chem.*, 54, 1982, pp. 67A-76A

[21] Jackson J E, Quality control methods for two related variables. *Industrial Quality Control,* 7, pp. 2-6, 1956

[22] Kuselman I, Fajgelj A IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for a limited number of participants—chemical analyticallaboratories (IUPAC Technical Report) Pure Appl. Chem., Vol. 82, No. 5, pp. 1099–1135, 2010.

[23] Müller C H, Uhlig S, Estimation of variance components with high breakdown point and high efficiency; Biometrika; 88: Vol. 2, pp. 353-366, 2001.

[24] Rousseeuw P J, Verboven S. Computational Statistics & Data Analysis 40 (2002) 741 – 758

[25] Silverman B W, *Density Estimation*. Chapman and Hall, London (1986)

[26] Scott D W, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley. (1992)

[27] Sheather S J, Jones M C A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society series B, Vol 53 pp 683-690, 1991.

[28] Thompson M. (2000), Analyst 125, 385-386

[29] Thompson M., Ellison S.L.R., Wood R., "The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories" (IUPAC Technical Report), in *Pure and Applied Chemistry*, Vol. 78, No. 1, pp. 145-196, 2006

[30] Thompson M, Willetts P, Anderson S, Brereton P, Wood R; Collaborative trials of the sampling of two foodstuffs, wheat and green coffee; Analyst, 127, 689-691, 2002

[31] van Nuland Y, ISO 9002 and the circle technique, Qual. Eng., 5 1992, pp. 269-291

[32] Uhlig, S, Robust estimation of variance components with high breakdown point in the 1-way random effect model. In: Kitsos, C.P. und Edler, L.; Industrial Statistics; Physica, S. 65-73, 1997.