INTERNATIONAL STANDARD

ISO 13528

Second edition
2015-08-01

Corrected version
2016-10-15

Statistical methods for use in proficiency testing by interlaboratory comparison

Méthodes statistiques utilisées dans les essais d'aptitude par comparaison interlaboratoires

Reference number
ISO 13528:2015(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 6, *Measurement methods and results*.

This second edition of ISO 13528 cancels and replaces the first edition (ISO 13528:2005), of which it constitutes a technical revision. This second edition provides changes to bring the document into harmony with ISO/IEC 17043:2010, which replaced ISO Guide 43-1:1997. It follows a revised structure, to describe better the process of the design, analysis, and reporting of proficiency testing schemes. It also eliminates some procedures that are no longer considered to be appropriate, and adds or revises some other sections to be consistent with ISO/IEC 17043 and to provide clarity and correct minor errors. New sections have been added for qualitative data and additional robust statistical methods.

This corrected version of ISO 13528:2015 incorporates the following corrections:

— 7.5.2.2, equation (5):"U_{char}" has been replaced with "u_{char}";

— 9.8.2, NOTE: reference to "E.3" has been replaced with a reference to "E.4";

— 10.3.2, iv, equation (19): an "addition" sign has been added between "q_{min}"and "(i-1)";

— B.2.3, b), line before Table B.1: Reference "[33]" has been replaced with Reference "[32]";

— B.2.3, Table B.1, first row, first column: "*gm*" has been replaced with "g";

— B.2.3, paragraph below Table B.1: in the first formula,"F_{2m}" has been replaced with "F_m" and subscript for "-1" has been removed; in the second formula, subscript for "-1" has been removed;

— B.2.3 b), NOTE: the equation for F_1 is now divided by "(g-1)";

— B.3, equations (B.7), (B.14) and (B.16): the terms s_x^2 and s_x have been replaced with $s_{\bar{x}}^2$ and $s_{\bar{x}}$; in addition, in equation (B.16) the square root symbol has been moved outside the "max (0, ...)" bracket;

— B.3, equation (B.8): "s_t^2" has been replaced with "w_t^2";

— C.3.1, NOTE 2, first line: "the" has been removed before "identical";

— C.3.1, paragraph after equation (C.10), second line: the words "the modified data in" have been deleted.

— C.3.1, last note: "NOTE" has been replaced with "NOTE 3", and Reference to "E.3 and E.4" has been replaced with a Reference to "E.1 and E.3";

The following minor editorial corrections have been implemented for consistency throughout the document:

— 8.3.1, third bullet, last line: the first occurrence of "approved" (after "more") has been deleted;

— 8.6.1, first line: "$,\sigma_{pt},$" has been replaced with "(σ_{pt})" (for presentation consistency);

— B.4.1.2, second bullet, second line: the word "samples" has been replaced with "proficiency testing items" (for terminological consistency)

— Annexes D and E, titles: the first letters in all words after the first one is now in lower case (for presentation consistency).

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 proficiency testing 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, on the requirements for the statistical design, validation of proficiency test items, review of results, and reporting summary statistics. Annex B of ISO/IEC 17043:2010 briefly describes the general statistical methods that are used in proficiency testing schemes. This International Standard is intended to be complementary to ISO/IEC 17043, 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 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 incorporates published guidance for the proficiency testing of chemical analytical laboratories [32] but additionally includes a wider range of procedures to permit use with valid measurement methods and qualitative identifications. This revision of ISO 13528:2005 contains most of the statistical methods and guidance from the first edition, extended as necessary by the previously referenced documents and the extended scope of ISO/IEC 17043. ISO/IEC 17043 includes proficiency testing for individuals and inspection bodies, and Annex B, which includes considerations for qualitative results.

This Standard includes 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 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 and shall authorize specific personnel to conduct statistical analysis. Neither ISO/IEC 17043 nor this International Standard 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 have experience or training in the common techniques applicable to the analysis of data from proficiency testing schemes. 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 routines for a particular scheme or analysis; some techniques will require computer applications that are freely available (at the time of publication of this Standard). In all cases, the users 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 by interlaboratory comparison

1 Scope

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

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

This Standard is applicable to proficiency testing where the results reported are either quantitative measurements 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 proficiency test items into known categories by inspection - or in determining by inspection whether proficiency test items arise, or do not arise, from the same original source - and the classification results are compared objectively, the provisions of this Standard that relate to nominal (qualitative) properties may apply.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO Guide 30, *Reference materials — Selected terms and definitions*

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

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)*

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 30, and the following apply. In the case of differences between these references on the use of terms, definitions in ISO 3534 parts 1-2 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 1 to entry: For the purposes of this International Standard, the term "proficiency testing" is taken in its widest sense and includes, but is not limited to:

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

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

— 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;

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

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

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

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

— 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
value attributed to a particular property of a proficiency test item

3.4
standard deviation for proficiency assessment
measure of dispersion used in the evaluation of results of proficiency testing

Note 1 to entry: This can be interpreted as the population standard deviation of results from a hypothetical population of laboratories performing exactly in accordance with requirements.

Note 2 to entry: The standard deviation for proficiency assessment applies only to ratio and interval scale results.

Note 3 to entry: Not all proficiency testing schemes evaluate performance based on the dispersion of results.

[SOURCE: ISO/IEC 17043:2010, modified — In the definition ", based on the available information" has been deleted. Note 1 to the entry has been added, and Notes 2 and 3 have been slightly edited.]

3.5
measurement error
measured quantity value minus a reference quantity value

[SOURCE: ISO/IEC Guide 99:2007, modified — Notes have been deleted.]

3.6
maximum permissible error
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

[SOURCE: ISO/IEC Guide 99:2007, modified — Notes have been deleted.]

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

Note 1 to entry: A common variation on the z score, sometimes denoted z' (commonly pronounced z-prime), is formed by combining the uncertainty of the assigned value with the standard deviation for proficiency assessment before calculating the z score.

3.8
zeta score
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
proportion of allowed limit score
standardized measure of performance, calculated using the participant result, assigned value and the criterion for measurement error in a proficiency test

Note 1 to entry: For single results, performance can be expressed as the deviation from the assigned value (D or $D\ \%$).

3.10
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 an indication of a need to investigate possible causes; a z score in excess of 3 is conventionally taken as an action signal indicating a need for corrective action.

3.11
consensus value
value derived from a collection of results in an interlaboratory comparison

Note 1 to entry: 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 values derived from results of a specified subset of such results or, for example, from a number of expert laboratories.

3.12
outlier
member of a set of values which is inconsistent with other members of that set

Note 1 to entry: 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 blunder.

Note 2 to entry: 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.

[SOURCE: ISO 5725-1:1994, modified — The Notes to the entry have been added.]

3.13
participant
laboratory, organization, or individual that receives proficiency test items and submits results for review by the proficiency testing provider

3.14
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 1 to entry: In most instances, proficiency test items meet the ISO Guide 30 definition of "reference material" (3.17).

3.15
proficiency testing provider
organization which takes responsibility for all tasks in the development and operation of a proficiency testing scheme

3.16
proficiency testing scheme
proficiency testing designed and operated in one or more rounds for a specified area of testing, measurement, calibration or inspection

Note 1 to entry: 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.

3.17
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

Note 1 to entry: RM is a generic term.

Note 2 to entry: Properties can be quantitative or qualitative, e.g. identity of substances or species.

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

[SOURCE: ISO Guide 30:2015, modified —Note 4 has been deleted.]

3.18
certified reference material
CRM
reference material (RM) characterized by a metrologically valid procedure for one or more specified properties, accompanied by an RM certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability

Note 1 to entry: The concept of value includes a nominal property or a qualitative attributes such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities or levels of confidence.

[SOURCE: ISO Guide 30:2015, modified —Notes 2, 3 and 4 have been deleted.]

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. Any statistical assumptions on which the methods or design are based shall be stated in the design or in a written description of the proficiency testing scheme, and these assumptions shall be demonstrated to be reasonable.

NOTE A statistically valid method has a sound theoretical basis, has known performance under the expected conditions of use and relies on assumptions or conditions which can be shown to apply to the data sufficiently well for the purpose at hand.

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

4.1.3 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 of the proficiency testing scheme or in a separate summary of procedures that is available to participants.

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

4.2 Basic model

4.2.1 For quantitative results in proficiency testing schemes where a single result is reported for a given proficiency test item, the basic model is given in equation (1).

$$x_i = \mu + \varepsilon_i \tag{1}$$

where

x_i = proficiency test result from participant i

μ = true value for the measurand

ε_i = measurement error for participant i, distributed according to a relevant model

NOTE 1 Common models for ε include: the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$ with mean 0 and variance either constant or different for each laboratory; or more commonly, an 'outlier-contaminated normal' distribution consisting of a mixture of a normal distribution with a wider distribution representing the population of erroneous results.

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.

NOTE 3 This model differs from the basic model in ISO 5725, in that it does not include the laboratory bias term B_i. This is because the laboratory bias and residual error terms cannot be distinguished when only one observation is reported. Where a participant's results from several rounds or test items are considered, however, it may become useful to include a separate term for 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 determining the assigned value and for determining the criteria for performance evaluation; for example when the assigned value is the robust mean of participant results and the performance evaluation is derived from σ_{pt} or δ_E, where δ_E is a predefined allowance for measurement error and $\sigma_{pt} = \delta_E / 3$; 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 measurement uncertainty, the assigned value is typically an appropriate reference value.

5 Guidelines for the statistical design of proficiency testing schemes

5.1 Introduction to the statistical design of proficiency testing schemes

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 assessed 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 performance 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 measurement uncertainty of the difference. Examination of the performance scores over multiple 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, 4.4.4.1, the statistical design "shall be developed to meet the objectives of the proficiency testing 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.

Design considerations for common objectives are listed below. Other objectives are possible.

EXAMPLE 1 For 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, the design will require a method for obtaining an externally defined reference value, a method of setting limits, and a scoring method;

EXAMPLE 2 For a proficiency testing scheme to compare a participant's result with combined results from a group in the same round, and limits that are specified before the round begins, the design will need to consider how the assigned value will be determined from the combined results as well as methods for setting limits and scoring;

EXAMPLE 3 For a proficiency testing scheme to compare a participant's result with combined results from a group in the same round, and limits determined by the variability of participant results, the design will need to consider the calculation of an assigned value and an appropriate measure of dispersion as well as the method of scoring;

EXAMPLE 4 For a proficiency testing scheme to compare a participant's result with the assigned value, using the participant's own measurement uncertainty, the design will need to consider how the assigned value and its uncertainty are to be obtained and how participant measurement uncertainties are to be used in scoring.

EXAMPLE 5 For a proficiency testing scheme with an objective to compare the performance of different measurement methods, the design will need to consider the relevant summary statistics and procedures to calculate them.

5.2.2 There are various types of data used in proficiency testing, including quantitative, nominal (categorical), and ordinal. Among the quantitative variables, some results might be on an interval scale; or a relative, or ratio scale. For some measurements on a quantitative 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 quantitative variables.

NOTE 1 For quantitative values, an interval scale is a scale on which intervals (differences) are meaningful but ratios are not, such as the Celsius temperature scale. A ratio scale is a scale on which intervals and ratios are both meaningful, such as the Kelvin temperature scale, or most common units for length.

NOTE 2 For qualitative values, a categorical scale has distinct values for which ordering is not meaningful, such as the names of bacterial species. Values on an ordinal scale have a meaningful ordering but differences are not meaningful; for example a scale such as 'large, medium, small' can be ordered but the differences between values are undefined other than in terms of the number of intervening values.

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 shall be appropriate for all the stated purposes for the particular proficiency testing scheme.

5.3 Considerations for the statistical distribution of results

5.3.1 ISO/IEC 17043:2010, 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 competently determined 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.

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

5.3.1.2 When the distribution expected for the proficiency testing scheme is not sufficiently symmetric (allowing for 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 may be appropriate as the first step in analysis.

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

5.3.1.3 In some areas of calibration, participant results may follow statistical distributions that are described in the measurement procedure (for example, exponential, or a wave form); these defined distributions should be considered in any evaluation protocol.

5.3.2 According to ISO/IEC 17043:2010, 4.4.4.2, the proficiency testing provider shall state the basis for any statistical assumptions and demonstrate that the assumptions are reasonable. This demonstration may be based on, for example, the observed data, results from previous rounds of the proficiency testing 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 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, 4.4.4.3 b)). 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 The IUPAC/CITAC Technical Report: *Selection and use of proficiency testing schemes for a limited number of participants* [24] provides useful guidance for proficiency testing schemes where there are few participants. 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 of a proficiency testing scheme.

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

— the statistical methods used, for example the particular robust method or outlier removal strategy chosen;

— the experience of the participants with the particular proficiency testing scheme;

— the experience of the proficiency testing provider with the matrix, measurand, methods, and group of participants;

— whether the intent is to determine the assigned value or the standard deviation (or both).

Further guidance on techniques for handling a small number of participants is provided in Annex D.1.

5.5 Guidelines for choosing the reporting format

5.5.1 It is a requirement of ISO/IEC 17043:2010, 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 1 For some proficiency testing schemes, an objective is to evaluate a participant's ability to follow a standard method, which could include the use of a particular unit of measurement or number of significant digits.

NOTE 2 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 If a proficiency testing scheme requires replicate measurements on proficiency test items, the participant should be required to report all replicate values. This can occur, for example, if an objective is to evaluate a participant's precision on known replicate proficiency 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 value (or other estimate of location) and uncertainty to assist data analysis by the proficiency testing provider.

5.5.3 Where conventional reporting practice is to report results as 'less than' or 'greater than' a limit (such as a calibration level or a quantitation limit) and where numerical results are required for scoring, the proficiency testing provider shall determine how the results will be processed.

5.5.3.1 The proficiency testing provider should either adopt validated data treatment and scoring procedures that accommodate censored data (see Annex E.1), or require participants to report the measured value of the result either in place of, or in addition to, the conventional reported value.

NOTE 1 An option of scoring procedure could be to not score such data.

NOTE 2 Requiring participants to report numerical values outside the range normally reported (for example, below the participant's quantitation limit) will permit use of statistical methods that require numerical values but may result in scores that do not reflect the participant's routine service to customers.

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[21], 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 each of the alternative statistical procedures considered potentially applicable in the circumstances, and investigate the importance of any difference(s).

5.5.3.3 Where censored results such as 'less than' statements are expected or have been observed, the proficiency testing scheme design should include provisions for scoring and/or other action on censored values reported by participants, and participants should be notified of these provisions.

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

5.5.4 Usually, the number of significant digits to report will be determined by the design of the proficiency testing scheme.

5.5.4.1 When specifying numbers of significant digits to be reported, the rounding error should be negligible compared to the expected variation between participants.

NOTE 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.

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 proficiency testing provider may specify the number of digits to be reported.

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 proficiency testing provider could then require participants to report to two or three decimal places to obtain sufficiently reliable estimates of location and variation.

5.5.4.3 If it is allowed that different participants will report results using different numbers of significant digits, the proficiency testing provider should 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. The provider shall assess homogeneity and stability using criteria that 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 or more of the following approaches:

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

b) experience with the behaviour of closely similar proficiency test items in previous rounds of the proficiency testing scheme, verified as necessary for the current round;

c) assessment of participant data in the current round of the proficiency testing scheme for evidence of consistency with previous rounds, for evidence of change with reporting time or production order, or any unexpected dispersion attributable to inhomogeneity or instability.

NOTE 1 These approaches can be adopted on a case-by-case basis, using appropriate statistical techniques and technical justification. The approach will often change during the lifetime of a proficiency testing scheme, for example as accumulated experience reduces the initial requirement for experimental study.

NOTE 2 Relying on experience (as in b above) is only reasonable so long as:

1. The process for producing batches of the proficiency test item(s) does not change in any way that may impact homogeneity;

2. The materials used in production of the proficiency test item(s) do not change in any way that may impact homogeneity;

3. There is not a "failure" in homogeneity identified via either homogeneity testing or participant responses; and,

4. The homogeneity requirements for the material are reviewed regularly, taking account of the intended use of the material at the time of the review, to ensure that the homogeneity achieved by the production process remains fit for purpose.

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, there 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 of the proficiency testing scheme. This should include consideration of tendencies for particular proficiency test items 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 All measurands (or properties) should normally be checked for homogeneity and stability. However, where the behaviour of a subset of properties can be shown to provide a good indication of stability and/or homogeneity for all properties reported on in a round, the assessment described in section 6.1.1 may be limited to that subset of properties. The measurands that are checked should be sensitive to sources of inhomogeneity or instability in the processing of the proficiency test item. Some important cases 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.

EXAMPLE In a proficiency testing scheme for the toxic metal content of soils, measured metal content is primarily affected by moisture content. A check for consistent moisture content may then be considered sufficient to ensure adequate stability of toxic metals.

NOTE An example of homogeneity and stability checks is provided in Annex E.2, using statistical methods recommended in Annex B.

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 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 use 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 measurands, 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, 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.

NOTE Obvious blunders, such as reporting results in incorrect units or switching results from different proficiency test items, occur 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 As a first step in any data analysis the provider should arrange for 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 (section 10.2). 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 (section 10.3).

NOTE 2 Other review techniques can be useful, such as a cumulative distribution plot or a stem-and-leaf diagram. Some graphical methods for data review are illustrated in Annexes E.3 and E.4.

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 of the proficiency testing scheme.

6.5 Robust statistical methods

6.5.1 Robust statistical methods can be used to describe the central part of a normally distributed set of results, but without requiring the identification of specific values as outliers and excluding them from subsequent analyses. Many robust techniques used are based (in the first step) on the median and the range of the central 50 % of results - these are measures of the center and spread of the data, similar to the mean and standard deviation. In general, robust methods should be used in preference to methods that delete results labelled as outliers.

NOTE Strategies that apply classical statistics such as the standard deviation after removing outliers usually lead to under-estimates of dispersion for near-normal data; robust statistics are usually adjusted to give unbiased estimates of dispersion.

6.5.2 The median, scaled median absolute deviation (*MADe*), and normalized IQR (*nIQR*) are allowed as simple estimators. Algorithm A transforms the original data by a process called winsorisation to provide alternative estimators of mean and standard deviation for near-normal data and is most useful where the expected proportion of outliers is below 20 %. The Q_n and Q methods (described in Annex C) for estimating standard deviation are particularly useful for situations where a large proportion (>20 %) of results can be discrepant, or where data cannot be reliably reviewed by experts. Other methods described in Annex C also provide good performance when the expected proportion of extreme values is over 20 % (see Annex D).

NOTE The median, interquartile range and scaled median absolute deviation have larger variance than the mean and standard deviation when applied to approximately normally distributed data. More sophisticated robust estimators provide better performance for approximately normally distributed data while retaining much of the resistance to outlying results that is offered by the median and interquartile range.

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 Details for robust procedures are provided in Annex C. Annexes E.3 and E.4 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, coupled with outlier rejection, to provide a degree of resistance to extreme values when calculating summary statistics. Where outlier detection techniques are used, the assumptions underlying the test should be demonstrated to

apply sufficiently for the purpose of the proficiency testing scheme; in particular, many outlier tests assume underlying normality.

NOTE ISO 16269-4 [10] and ISO 5725-2 [1] provide several outlier identification procedures that are applicable to inter-laboratory data.

6.6.2 Outlier rejection strategies, which are based on rejection of outliers detected by an outlier test at a high level of confidence, followed by application of simple statistics such as the mean and standard deviation, are permitted where robust methods are not applicable (see 6.5.1). Where outlier rejection strategies are used, the proficiency testing provider shall

a) document the tests and level of confidence required for rejection;

b) set limits for the proportion of data rejected by successive outlier tests, if used;

c) demonstrate that the resulting estimates of location and (if appropriate) scale have sufficient performance (including efficiency and bias) for the purposes of the proficiency testing scheme.

NOTE ISO 5725-2 provides recommendations for the level of confidence appropriate for outlier rejection in interlaboratory studies for the determination of precision of test methods. In particular, ISO 5725-2 recommends rejection only at the 99 % level unless there is other strong reason to reject a particular result.

6.6.3 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 proficiency testing scheme.

NOTE 1 Outliers among reported values are often identified by employing the 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 the 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 of a proficiency testing scheme, it is common to use Cochran's test for repeatability outliers, also described in ISO 5725-2.

NOTE 3 Outliers can 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[13].

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 a proficiency testing scheme. This is discussed in section 7.8.

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

7.1.4 The method of determining the assigned value and its associated uncertainty shall be stated in each report to participants or clearly described in a scheme protocol available to all participants.

7.2 Determining the uncertainty of the assigned value

7.2.1 The *Guide to the expression of uncertainty in measurement* (ISO/IEC Guide 98-3[14]) 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:

$$x_{pt} = x_{char} + \delta_{hom} + \delta_{trans} + \delta_{stab} \quad (2)$$

where

x_{pt} denotes the assigned value;

x_{char} denotes the property value obtained from the characterization (determination of assigned value);

δ_{hom} denotes an error term due to the difference between proficiency test items;

δ_{trans} denotes an error term due to instability under transport conditions;

δ_{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:

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

where

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

u_{char} denotes the standard uncertainty due to characterization;

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

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

u_{stab} denotes the 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 in large part reflected in the variability of participant results. In this case the uncertainty of characterization, as described in sections 7.3-7.7, is 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 δ_{trans} and δ_{stab} are negligible. 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 shall be included in the evaluation of the uncertainty of the assigned value.

7.3 Formulation

7.3.1 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.

7.3.1.1 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 is similar to customer samples that are 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 proficiency test items in which the addition is more loosely bonded than in routinely tested samples, or in a different form, it may be preferable to use another approach to prepare the proficiency test items.

7.3.1.4 Determination of the assigned value by formulation is 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 (see section 7.5).

7.3.2 When the assigned value is calculated from the formulation of the proficiency test item, the standard uncertainty for the characterization (u_{char}) is estimated by combination of uncertainties using an appropriate model. For example, in proficiency testing for chemical measurements the uncertainties will usually be those associated with gravimetric and volumetric measurements and the purity of any materials used in formulation. 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 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 that:

— 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 proficiency test items.

— a CRM may be known to the participants making it important to conceal the identity of the proficiency test item.

7.4.2 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 An assigned value can be determined by a single laboratory using a reference method, such as, for example, a primary method. The reference method used should be completely described and understood, and with a complete uncertainty statement and documented metrological traceability that is appropriate for the proficiency testing scheme. The reference method should be commutable for all measurement methods used by participants.

7.5.1.1 The assigned value should be the average from a designed study using more than one proficiency test item or measurement conditions, and a sufficient number of replicate measurements.

7.5.1.2 The uncertainty of characterization is the appropriate estimate of uncertainty for the reference method and the designed study conditions.

7.5.2 The assigned value x_{pt} of the proficiency test item can be derived by a single laboratory using a suitable measurement method, from a calibration against the reference values of a closely matched certified reference material. This approach assumes that the CRM is commutable for all measurement methods used by participants.

7.5.2.1 This determination requires a series of tests to be carried out, in one laboratory, on proficiency test items and the CRM, using the same measurement method, and under repeatability conditions. When

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,

$$x_{pt} = x_{CRM} + \bar{d} \tag{4}$$

NOTE x_{CRM} and $\bar{d}$ are independent except in the rare situation that the expert laboratory also produced the CRM.

7.5.2.2 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 (5).

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

The example in Annex E.5 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.

7.5.3 When a reference value is assigned prior to the commencement of a round of a sequential proficiency testing scheme, and then the reference value is subsequently checked using the same measuring system, the difference between the values shall be less than two times the uncertainty of that difference (that is, the results shall be metrologically compatible). In such cases the proficiency testing provider may choose to use an average of the measurements as the assigned value, with the appropriate uncertainty. If the results are not metrologically compatible, the proficiency testing provider should investigate the reason for the difference, and take appropriate steps, including use of alternative methods to determine the assigned value and its uncertainty or abandonment of the round.

7.6 Consensus value from expert laboratories

7.6.1 Assigned values can be determined using an interlaboratory comparison study with expert laboratories, as described 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 proficiency test items are then selected at random and analysed by a group of experts using a protocol that specifies the numbers of proficiency test items and replicates and any other relevant conditions. Each expert laboratory is required to provide a standard uncertainty with their results.

7.6.2 Where the expert laboratories report a single result and are not required by the measurement protocol to provide sufficient uncertainty information with results, or where evidence from the reported results or elsewhere suggests that the reported uncertainties are not sufficiently reliable, the consensus value should normally be obtained by the methods of section 7.7, applied to the set of expert laboratory results. Where the expert laboratories report more than one result each (for example, including replicates), the proficiency testing scheme provider shall establish an alternative method of determining the assigned value and associated uncertainty that is statistically valid (see 4.1.1) and allows for the possibility of outliers or other departures from the expected distribution of results.

7.6.3 Where the expert laboratories report uncertainties with the results, the estimation of a value by consensus of results is a complex problem and a wide variety of approaches has been suggested, including, for example, weighted averages, un-weighted averages, procedures that make allowance for over dispersion and procedures that allow for possible outlying or erroneous results and uncertainty estimates[16]. The proficiency testing provider shall accordingly establish a procedure for estimation that:

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

b) should use a weighting procedure appropriate for the scale and reliability of the reported uncertainties, which may include equal weighting if the reported uncertainties are either similar or of poor or unknown reliability (see 7.6.2);

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

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

e) should have a sound theoretical basis;

f) shall have demonstrated performance (for example on test data or in simulations) sufficient for the purposes of the proficiency testing scheme.

7.7 Consensus value from participant results

7.7.1 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) formed from 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.

7.7.1.1 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.

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

7.7.1.3 The advantages of this approach are that:

a) no additional measurements are required to obtain the assigned value;

b) the approach may be particularly useful with a standardized, operationally defined measurand, as there is often no more reliable method to obtain equivalent results.

7.7.1.4 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 and control of other relevant method conditions by all of the participants contributing to the consensus value.

7.7.2 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 [17,18] give details of bootstrapping techniques.

NOTE An example using a bootstrap technique is provided in Annex E.6.

7.7.3 When the assigned value is derived as a robust average calculated using procedures in Annex C.2, C.3, or C.5, the standard uncertainty of the assigned value x_{pt} may be estimated as:

$$u\left(x_{pt}\right) = 1,25 \times \frac{s^*}{\sqrt{p}} \tag{6}$$

where s^* is the robust standard deviation of the results. (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.

NOTE 3 An example of using an assigned value from participant results is provided in Annex 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 is estimated as:

$$u_{diff} = \sqrt{u^2\left(x_{ref}\right) + u^2\left(x_{pt}\right)} \tag{7}$$

where

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

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

NOTE An example of a comparison of a reference value with a consensus value is included in Annex E.7.

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

— bias in the reference 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 difference, the proficiency testing provider should decide whether to evaluate results or not, and (for continuous proficiency testing schemes), whether to amend the design for subsequent proficiency testing 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 proficiency testing schemes.

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', ζ, E_n), as discussed in sections 9.4-9.7. This can also be done by comparing the difference with a defined criterion (D or $D\%$ compared to δ_E) as discussed in 9.3. An alternative approach to evaluation is to compare the difference with a participant's claim for the uncertainty of their result combined with the uncertainty of the assigned value (E_n and ζ).

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 may be divided by the action limit to obtain σ_{pt}. A prescribed maximum permissible error may be used directly as δ_E for use with D or $D\%$. The advantages of this approach for continuous schemes are:

a) performance scores have a consistent interpretation in terms of fitness for purpose from one round to the next;

b) performance scores are not subject to the variation expected when estimating dispersion from reported results.

EXAMPLE If a regulatory criterion is specified as a maximum permissible 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 of the proficiency testing scheme, 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 performance 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 maximum permissible 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 maximum permissible 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.

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 maximum permissible 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 of the proficiency testing scheme 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 proficiency testing providers approved for an area of testing or calibration.

8.3.2 The review of previous rounds of a proficiency testing scheme 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[31]. Standard deviations and relative standard deviations should be considered, with selection based on which is more consistent across the appropriate range of measurand levels. Appropriate maximum permissible error can also be obtained in this manner.

8.3.3 When the criterion for evaluation of performance is based on consensus statistics from previous rounds of a proficiency testing scheme, robust estimates of the standard deviation should be used.

NOTE 1 Algorithm S (Annex C.4) provides a robust pooled standard deviation that is applicable when all previous rounds of a proficiency testing scheme under consideration have the same expected standard deviation or (if relative deviations are used for the assessment) the same relative standard deviation.

NOTE 2 An example of deriving a value from experience of previous rounds of a proficiency testing scheme is provided in Annex E.8.

8.4 By use of a general model

8.4.1 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.

8.4.2 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.

8.4.3 A specific estimation taking the specificities of the measurement problem into consideration is generally preferable to a generic approach. Consequently, before using a general model, the possibility of using the approaches described in 8.2, 8.3 and 8.5 should be explored.

EXAMPLE Horwitz curve.

One common general model for chemical applications was described by Horwitz[22] and modified by Thompson[31]. 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:

$$\sigma_R = \begin{cases} 0,22c & \text{when} \quad c < 1,2\times10^{-7} \\ 0,02c^{0,8495} & \text{when} \quad 1,2\times10^{-7} \le c \le 0,138 \\ 0,01c^{0,5} & \text{when} \quad c > 0,138 \end{cases} \tag{8}$$

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

NOTE 1 The Horwitz model is empirical, based on observations from collaborative trials of many parameters over an extended time period. The σ_R values are the expected upper limits of interlaboratory variability when the collaborative trial had no significant problems. The σ_R values therefore might not be appropriate criteria for determining competence in a proficiency testing scheme.

NOTE 2 An example of deriving a value from the modified Horwitz model is provided in Annex E.9.

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:

$$\sigma_{pt} = \sqrt{\sigma_R^2 - \sigma_r^2\left(1 - 1/m\right)} \tag{9}$$

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

NOTE This equation is derived from a basic random effects model from ISO 5725-2.

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 an equivalent procedure.

NOTE An example is presented in Annex E.10.

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 calculated from the results of participants in the same round of the proficiency testing scheme. When this approach is used it is usually most convenient to use a performance score such as the z score. A robust estimate of the standard deviation of the results reported by all the participants, calculated using a technique listed in Annex C, should normally be used to calculate σ_{pt}. 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 (section 9.3.6).

8.6.2 The use of participant results can lead to criteria for performance evaluation that are not appropriate. The proficiency testing provider should ensure that the σ_{pt} used for performance evaluations is fit for purpose.

8.6.2.1 The proficiency testing provider should place a limit on the lowest value of σ_{pt} that will be used, in the case that the robust standard deviation is very small. This limit should be chosen so that when measurement error is fit for the most challenging intended use, the performance score will be z<3,0.

EXAMPLE In a proficiency testing scheme for fabric, one measurand is number of threads per centimeter. The robust standard deviation can be small in some rounds (<1 thread per cm.), and errors less than 4 threads/cm are considered to be insignificant. The proficiency testing provider determines that the robust standard deviation is used as σ_{pt}, unless it is less than 1,3 threads/cm, in which case σ_{pt} =1,3 is used.

8.6.2.2 The proficiency testing provider should place a limit on the largest σ_{pt} that will be used, or on the measurement results that can be evaluated as “acceptable” (no signal), in the case that the robust standard deviation is very large. This limit should be chosen so that results that are not fit for purpose will receive an action signal.

8.6.2.3 In some cases the proficiency testing provider may place upper or lower limits on the interval of results that can be evaluated as ‘acceptable’ (no warning or action signal), when symmetric intervals include results that would not be fit for purpose.

EXAMPLE For a regulatory proficiency testing scheme for non-potable water, regulations specify that results must be within $3\sigma_{pt}$ of the robust mean of participant results. However, because in some cases the range of acceptable results could include 0 µg/L, any result less than 10 % of a formulated value shall generate an action signal (or 'unacceptable'). A proficiency testing item is formulated with 4,0 µg/L of a regulated substance. The robust participant mean is 3,2 µg/L and σ_{pt} is 1,1 µg/L. Therefore it is possible for a participant to submit a result of 0,0 µg/L and be within $3\sigma_{pt}$, but any result less than 0,4 µg/L will be evaluated as "unacceptable".

8.6.3 The main advantages of this approach are simplicity and conventional acceptance due to successful use in many situations. This may be the only feasible approach.

8.6.4 There are several disadvantages with this approach:

a) The value of σ_{pt} may vary substantially from round to round of a proficiency testing scheme, 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 of a proficiency testing scheme to the next.

c) Using dispersion measures derived from the data can 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.

NOTE Examples of using participant data are provided in the comprehensive example in Annex E.3.

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, 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 round of proficiency testing are greater by a factor of two or more from the values expected from prior data or experience, then the proficiency testing provider should investigate why agreement in this round 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 Performance scores should be easily reviewed across measurand levels and different rounds of a proficiency testing scheme.

9.1.3 Participant results should be reviewed and determined to be consistent with the assumptions used in the design of the proficiency testing scheme, to allow for meaningful performance statistics. For example, that 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 may be considered to be negligible and need not be included in the interpretation of the results of the round of proficiency testing.

$$u(x_{pt}) < 0{,}3\sigma_{pt} \quad \text{or} \quad u(x_{pt}) < 0{,}1\delta_E \tag{10}$$

NOTE $0{,}3\sigma_{pt}$ is equivalent to $0{,}1\delta_E$ when $|z| \geq 3{,}0$ generates an action signal.

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 (10).

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 ζ 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 [32] 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.

NOTE The techniques presented in this section are demonstrated in Annexes E.3 and E.4.

9.3 Estimates of deviation (measurement error)

9.3.1 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}:

$$D_i = x_i - x_{pt} \tag{11}$$

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:

$$D_i\% = 100\left(x_i - x_{pt}\right) / x_{pt}\% \tag{12}$$

9.3.2 The difference D or $D\%$ is usually compared with a criterion δ_E based on fitness for purpose or with experience from previous rounds of a proficiency testing scheme; the criterion is noted here as δ_E, an allowance for measurement error. If $-\delta_E < D < \delta_E$ then the performance is considered to be 'acceptable' (or 'no signal'). (The same criterion applies for $D\%$, depending on the expression of δ_E.)

9.3.3 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.

9.3.4 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. The advantage of $D\%$ is that understanding is intuitive, it is standardized for measurand level, and it is related to common causes of error (for example, incorrect calibration or bias in dilution).

9.3.5 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 $-\delta_E < D < \delta_E$.

9.3.6 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 a standardized performance score that shows 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:

$$P_{Ai} = (D_i / \delta_E) \times 100\ \% \tag{13}$$

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

NOTE 1 P_A scores can be compared across levels and different rounds of a proficiency testing scheme, or tracked in charts. These performance 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.

NOTE 2 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.

NOTE 3 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 The *z* score for a proficiency test result x_i is calculated as:

$$z_i = \frac{\left(x_i - x_{pt}\right)}{\sigma_{pt}} \quad (14)$$

where

x_{pt} is the assigned value, and

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

9.4.2 The conventional interpretation of *z* scores is as follows (see ISO/IEC 17043:2010, B.4.1.1):

— 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 $-3{,}0 \leq z \leq 3{,}0$ and only about 5 % would be expected to fall outside the range $-2{,}0 \leq z \leq 2{,}0$. 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 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.

9.4.3 The proficiency testing provider shall determine appropriate rounding for reported *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 rules for rounding shall be included in the information available to participants.

NOTE It is rarely useful to have more than two digits after the decimal for z scores.

9.4.4 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 is only a small number of participants, there may be no action signal given. In this case, graphical methods that combine performance 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 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):

$$z'_i = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2\left(x_{pt}\right)}} \tag{15}$$

NOTE When x_{pt} and/or σ_{pt} are calculated from participant results, the performance score is correlated with individual participant results, because individual results 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, performance scores including the uncertainty of the assigned value without including an allowance for correlation represent under-estimates of the scores that would result if the covariance were included. For example, when $u(x_{pt})=0{,}3\sigma_{pt}$ then there is an underestimate of about 10 % of the z' score. Therefore equation (15) can be used when x_{pt} and/or σ_{pt} are determined from participant results.

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

$$\delta_E' = \sqrt{\delta_E^2 + U^2\left(x_{pt}\right)} \tag{16}$$

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

9.5.3 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 and $D\%$ scores would then be compared with δ_E' (see 9.3).

9.5.4 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

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

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

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

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 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:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}} \tag{17}$$

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}.

NOTE 1 When the assigned value x_{pt} is calculated as the consensus value from participant results, 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, performance scores including the uncertainty of the assigned value without including an allowance for correlation represent under-estimates of the scores 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 performance scores. Equation (17) can therefore be used with consensus statistics without adjustment for correlation.

NOTE 2 ζ scores differ from E_n scores (section 9.7) 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.

9.6.2 Using ζ scores allows direct assessment whether laboratories are able to deliver correct results, i.e. results that agree with x_{pt} within their measurement uncertainties. ζ scores may be interpreted using the same critical values of 2,0 and 3,0 as for z scores, or with multiples from the participant's coverage factor used when estimating expanded uncertainty. However, an adverse ζ score may indicate either a large deviation of x_i from x_{pt}, an under-estimate of uncertainty on the part of the participant, or a combination of both.

NOTE It may be useful for the proficiency testing provider to give additional information about the validity of reported uncertainties. Useful guidelines for such assessment are suggested in section 9.8.

9.6.3 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.

NOTE 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 participant'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-x_{pt}) or the combined standard uncertainties in comparison with a target uncertainty.

9.7 E_n scores

9.7.1 E_n scores can be useful when an objective for the proficiency testing scheme is to evaluate a participant's ability to have results close to the assigned value within their claimed expanded 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:

$$\left(E_n\right)_i = \frac{x_i - x_{pt}}{\sqrt{U^2\left(x_i\right) + U^2\left(x_{pt}\right)}} \tag{18}$$

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 98-3 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 E_n scores should be interpreted with caution, because they are ratios 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_i-x_{pt}) is smaller than needed by the participant's customers.

NOTE While the interpretation of E_n scores can be difficult, that does 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 Evaluation of participant uncertainties in testing

9.8.1 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 ζ scores described in section 9.6, and E_n scores in section 9.7, are options for evaluation of results against the claimed uncertainty.

9.8.2 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 their evaluation of uncertainty is not counting all relevant components, or is over-counting some components;

c) proficiency testing can be used to confirm claims of uncertainty, and this is easiest when the uncertainty is reported with the result.

NOTE An example of the analysis of data when uncertainties are reported is in Annex E.4.

9.8.3 Where x_{pt} is determined using procedures in sections 7.3-7.6 and $u(x_{pt})$ 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 from participant results (section 7.7), then the proficiency testing provider should determine 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}.

9.8.4 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}.

NOTE The factor 1,5 is the upper limit of the variability in standard deviations that can be expected for a consensus standard deviation with 10 or more results, based on the square root of percentiles of the F distribution. Any proficiency testing provider adopting this procedure may wish to use a different multiplier.

9.8.5 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.

9.8.6 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 5th percentile and above the 95th percentile of the reported standard or expanded uncertainties);

— limits based on an assumed distribution with scale based on the dispersion of reported uncertainties;

— 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 only when the distribution is approximately normal.

9.9 Combined performance scores

9.9.1 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.

9.9.2 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.

9.9.3 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 performance 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.

9.9.4 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 evaluate performance.

9.9.5 Combined performance 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 performance 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.

10 Graphical methods for describing performance scores

10.1 Application of graphical methods

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', ζ, 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 accrediting body, to enable them to judge the overall effectiveness of the proficiency testing 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 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 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 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 1 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.

NOTE 2 An example of a histogram plot is provided in Annex E.3.

10.3 Kernel density plots

10.3.1 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 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 performance 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[30] which recommends s^* based on the normalised interquartile range ($nIQR$). Other bandwidth selection rules that provide similar results include that of Scott[29], which replaces the multiplier of 0,9 with 1,06. Reference [29] 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[32].

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

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_{nk} from

$$q_i = q_{\min} + (i-1)\frac{\left(q_{n_k} - q_1\right)}{n_k - 1} \tag{19}$$

v) Calculate n_k densities h_1 to $h_{n\mathrm{k}}$ from

$$h_i = \frac{1}{p}\sum_{j=1}^{p} \varphi\left(\frac{x_j - q_i}{\sigma_k}\right) \text{ for } i = 1 \text{ to } i = n_k \tag{20}$$

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

vi) Plot h_i against q_i .

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 spreadsheet 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.

NOTE 3 Examples of kernel density plots are given in Annexes E.3, E.4, and E.6.

10.3.3 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 performance 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 performance scores into a bar-plot as shown in Figure E.10, in which scores for each participant are grouped together. Other standardized performance 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).

NOTE An example of a bar plot with z scores is provided in Annex E.11.

10.5 Youden Plot

10.5.1 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 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.

NOTE 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.

10.5.3 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 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 participant bias that affects different proficiency test items in a similar way. If there is no apparent visual relationship between results (e.g., points are distributed approximately evenly in a circular region, usually with higher density towards the centre) than the measurement errors for the two proficiency test items are largely independent. 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 1 In studies where all participants use the same measurement method, or plots of results are from a single measurement method, if results lie along a line, this may be 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.

NOTE 2 An example of a Youden plot is provided in Annex E.12.

10.6 Plots of repeatability standard deviations

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

10.6.2 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

x^* = the robust average of $x_1, x_2, \ldots, x_p$, as calculated by Algorithm A

w^* = the robust pooled average of $s_1, s_2, \ldots, s_p$, 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

$$\left(\sqrt{m}\,\frac{x_i - x^*}{w^*}\right)^2 + \left(\sqrt{2(m-1)}\,\ln\left(\frac{s_i}{w^*}\right)\right)^2 \tag{21}$$

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

$$s = w^* \exp\left\{\pm \frac{1}{\sqrt{2(m-1)}} \sqrt{\chi^2_{2;0,99} - \left(\sqrt{m}\,\frac{x - x^*}{w^*}\right)^2}\right\} \tag{22}$$

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

$$x = x^* - w^*\sqrt{\frac{\chi^2_{2;0,99}}{m}} \quad \text{to} \quad x^* + w^*\sqrt{\frac{\chi^2_{2;0,99}}{m}} \tag{23}$$

NOTE This procedure is based on the Circle Technique introduced by van Nuland[36]. 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 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 difficult.

NOTE An example of a plot of repeatability standard deviations is provided in Annex E.13.

10.7 Split samples

10.7.1 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 1 This type of study is common, but often named differently, such as “paired sample” or “bilateral comparisons”.

NOTE 2 This split sample design should not be confused with the ‘split level’ design used in ISO 5725, which involves two test items with slightly different levels supplied to all participants.

10.7.2 The data from a split-sample design 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 proficiency test item. Bivariate plots using the full range of concentrations can have a scale that makes it difficult to identify 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.

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

10.8.1 When standardized performance scores are to be combined over several rounds of a proficiency testing scheme, 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 performance scores for several rounds of a proficiency testing scheme are combined, can allow trends, and other features of the results, to be identified that are not apparent when performance scores for each round are examined separately.

NOTE With the use of "running scores" or "cumulative scores", in which the performance scores obtained by a participant are combined over several rounds of a proficiency testing scheme, the performance scores should be displayed graphically. 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. However, in some circumstances (e.g. with frequent rounds) 'smoothing' of occasional outlying scores may be helpful in demonstrating the underlying performance more clearly.

10.8.2 The Shewhart control chart is an effective method of identifying problems that cause large erratic values of z scores. See ISO 7870-2[6] for advice on plotting Shewhart charts and rules for action limits.

10.8.2.1 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. When several characteristics are measured in each round, the performance 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 of the proficiency testing scheme the performance 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.

10.8.2.2 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 (± 3,0 for z scores, or 100 % for P_A);

b) two out of three successive points outside either warning limit (± 2,0 for z scores or 70 % for P_A);

c) six consecutive results either positive or negative.

10.8.2.3 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 $(x_i - x_{pt})$, so the probability levels that are usually associated with the action and warning limits of a Shewhart control chart may not apply.

10.8.3 When the level of a property varies from one round of a proficiency testing scheme to another, plots of standardized performance 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 performance scores can all be plotted independently.

NOTE 1 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.

NOTE 2 An example of such a plot is provided in Annex E.14, 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; and

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

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

NOTE Guidance for ordinal data does not apply to measurement results that are based on a quantitative scale with discontinuous indications (such as dilutions or titres), see section 5.2.2.

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 proficiency test items. The proficiency testing provider may choose to circulate proficiency 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 two or more 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 two or more proficiency test items may provide additional

information on the nature of errors and also allow more sophisticated scoring of proficiency testing performance.

EXAMPLE 1 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 provider 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 performance scores to be allocated to participants on the basis of estimated probabilities of particular patterns of response.

EXAMPLE 2 Proficiency testing in forensic comparisons often requires matching proficiency test items as to whether they came from the same source or different sources (for example, fingerprints, DNA, bullet shell casings, footprints, etc.). In many cases "indeterminate" is an allowed response. A proficiency testing scheme might include multiple proficiency test items from different sources, and participants are asked to state which are from "same source", "different source", or "indeterminate" for every pair. This allows objective scores of number (or %) correct or incorrect, or number (%) correct matches, or correct rejections. Performance criteria can then be determined on fitness for use, or on degree of difficulty of the challenge.

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.

11.3 Assigned values for qualitative proficiency testing schemes

11.3.1 Values may be assigned to proficiency test items:

a) by expert judgement;

b) by use of reference materials as proficiency test items;

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

d) using the mode or median of participant results (the median is appropriate only for ordinal values).

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 in qualitative proficiency testing schemes. 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 section 11.3.1. If that is not appropriate the proficiency test item should not be used for performance assessment of participants.

NOTE In some cases it is possible for a single expert to determine the assigned value.

11.3.3 Where a reference material is provided to participants as a proficiency test item, the associated reference value, 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 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 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 proficiency testing 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 proficiency test item. For example, a proficiency test item 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 or doubts about origin that may compromise use of the proficiency test item.

NOTE Further detail on characterisation of such proficiency test items 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 proficiency testing 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 performance 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. Such scoring designs should assign lower scores to better performance, to be consistent with other types of performance scores (for example, z score, P_A score, C, and E_n).

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. 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 0 and the score is increased by 2 for each difference in rank until the score increases to a maximum of 6 (so a result adjacent to the assigned value would attract a score of 2).

Individual performance scores for each proficiency test item should be provided to participants. Where replicate observations are performed a summary of performance scores for each result 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 performance scores or score summaries in performance assessment. Combined performance scores or summaries may be calculated as, for example:

— the simple sum of performance scores across all proficiency test items;

— the count of each level of performance 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[20]. 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 performance 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.

NOTE An example of the analysis of ordinal data is provided in Annex E.15.

Annex A
(normative)

Symbols

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 permissible error criterion for differences
δ_{hom}	Error due to the difference between proficiency test items
δ_{stab}	Error due to instability during the period of proficiency testing
δ_{trans}	Error due to instability under transport conditions
E_n	"Error, normalized" score that includes uncertainties for the participant result and the assigned value
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
P_A	Proportion of allowed error (D/δ_E), can be expressed as a percentage
s_r	Estimate of repeatability standard deviation
s_R	Estimate of reproducibility standard deviation
s_s	Estimate of between-sample standard deviation
s^*	Robust estimate of the participant standard deviation
$s_{\bar{x}}$	Standard deviation of sample averages
s_w	Within-sample or within-laboratory standard deviation
σ_k	Bandwidth standard deviation used for kernel density plots
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
u_{hom}	Standard uncertainty due to the difference between proficiency test items
u_{stab}	Standard uncertainty due to instability during the period of proficiency testing
u_{trans}	Standard uncertainty due to instability under transport conditions
$u(x_i)$	Standard uncertainty of a result from participant i
$u(x_{pt})$	Standard uncertainty of the assigned value
$u(x_{ref})$	Standard uncertainty of a reference value
$U(x_i)$	Expanded uncertainty of reported result from participant i
$U(x_{pt})$	Expanded uncertainty of the assigned value
$U(x_{ref})$	Expanded uncertainty of a reference value
w_t	Between-test-portion range
w^*	Robust estimate of participant repeatability
x	Measurement result (generic)
x_{char}	Property value obtained from the determination of the assigned value
x_{CRM}	Assigned value for a property in a Certified Reference Material

x_i	Measurement result from participant *i*
x_{pt}	Assigned value
x_{ref}	Reference value for a stated purpose
x^*	Robust estimate of the participant mean
$\overline{x}$	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	Zeta score – modified *z* score that includes uncertainties for the participant result and the assigned value

Annex B
(normative)

Homogeneity and stability of proficiency test items

B.1 General procedure for a homogeneity check

B.1.1 To conduct an assessment for homogeneity for a bulk preparation of proficiency test items, follow the procedure given below:

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

Choose a laboratory to carry out the homogeneity check and a measurement method to use. The method should have a sufficiently small repeatability standard deviation (s_r) 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.

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

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 proficiency test items included in the homogeneity check may be reduced if suitable data are available from previous homogeneity checks on similar proficiency test items prepared by the same procedures.

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

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

Calculate the general average $\bar{x}$, within-sample standard deviation s_w, and between-sample 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 a sufficiently low repeatability standard deviation 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, take appropriate corrective action regarding the measurement method, or use caution in the interpretation of the results.

b) Examine the results for proficiency test item averages by 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 the outlying group from the analysis or, if m>2 and the high variance is caused by a single outlier, remove the outlying point.

NOTE If m>2 and a single observation is removed, subsequent calculation of s_w and s_s will need to take the resulting imbalance into account.

B.2.2 Compare the between-sample 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:

$$s_s \leq 0{,}3\ \sigma_{pt} \quad \text{(B.1)}$$

NOTE 1 The justification for the factor of 0,3 is that when this criterion is met the between-sample standard deviation contributes less than 10 % of the variance for evaluation of performance, so the performance evaluation is unlikely to be affected.

NOTE 2 Equivalently, s_s can be compared to δ_E :

$$s_s \leq 0{,}1\delta_E \quad \text{(B.2)}$$

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-sample standard deviation as calculated in section B.3 and

F_1 and F_2 are from standard statistical tables, reproduced in Table B.1, for the number of proficiency test items selected and with each item tested in duplicate[32].

Table B.1 — Factors *F1* and *F2* 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

Where m>2, F_2 in B.2.3 b) and Table B.1 shall be replaced with $F_m = (F_{g-1,\, g(m-1),\, 0{,}95} - 1)/m$ where $F_{g-1,\, g(m-1),\, 0{,}95} - 1$ is the value exceeded with probability 0,05 by a random variable with an F-distribution with $g - 1$ and $g\,(m - 1)$ degrees of freedom.

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

$F_1 = \chi^2_{0{,}95(g-1)}/(g-1)$ where $\chi^2_{0{,}95(g-1)}$ is the value exceeded with probability 0,05 by a chi-squared random variable with $g - 1$ degrees of freedom, and

$F_2 = (F_{0{,}95\,(g-1;g)}\text{-}1)/2$ where $F_{0{,}95(g\text{-}1;g)}$ is the value exceeded with probability 0,05 by a random variable with an F-distribution with $g - 1$ and g degrees of freedom.

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

B.2.4 When σ_{pt} is not known in advance, for example when σ_{pt} is the robust standard deviation of participant results, the proficiency testing provider should choose other criteria for determining sufficient homogeneity. Such procedures could include:

a) check for statistically significant differences between proficiency test items using, for example, the Analysis of Variance F test at α=0,05 ;

b) use information from previous rounds of the proficiency testing scheme to estimate σ_{pt} ;

c) use data from a precision experiment (such as a reproducibility standard deviation as described in ISO 5725-2);

d) accept the risk of distributing proficiency test items that are not sufficiently homogeneous, and check the criterion after the consensus σ_{pt} has been calculated.

B.2.5 If the criteria for sufficient homogeneity are not met, the proficiency testing provider shall consider adopting one of the following actions.

a) Include the between-sample 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.

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

b) Include s_s in the uncertainty of the assigned value and use z' or δ_E' to assess performance (see 9.5);

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

The estimate of within-sample standard deviation s_w and between-sample standard deviation s_s may be calculated using analysis of variance as shown below. The method shown is for a chosen number g of proficiency test items, measured in replicate m times.

The data from a homogeneity check are represented by $x_{t,k}$

where

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

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

Define the proficiency test item average and variance as:

$$\bar{x}_t = \frac{1}{m}\sum_{k=1}^{m} x_k$$

$$s_t^2 = \frac{1}{m}\sum_{k=1}^{m}\left(x_k - \bar{x}_t\right)^2 \tag{B.4}$$

and the estimate of between-test-portion variance as:

$$w_t^2 = \frac{1}{\left(m-1\right)}\sum_{k=1}^{m}\left(x_k - \bar{x}_t\right)^2 \tag{B.5}$$

Calculate the general average:

$$\bar{\bar{x}} = \frac{1}{g}\sum_{t=1}^{g}\bar{x}_t \tag{B.6}$$

the estimate of the variance of sample averages:

$$s_{\bar{x}}^2 = \frac{1}{\left(g-1\right)}\sum_{t=1}^{g}\left(\bar{x}_t - \bar{\bar{x}}\right)^2 \tag{B.7}$$

and the within-sample variance:

$$s_w^2 = \frac{1}{g}\sum_{t=1}^{g} w_t^2 \tag{B.8}$$

Estimate the combined variance of s_s and s_w

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

Finally, estimate the between-sample variance as

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

NOTE In the case that $s_s^2 < 0$, then it is appropriate to use $s_s = 0$.

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

Define the sample averages as:

$$\bar{x}_t = \left(x_{t,1} + x_{t,2}\right) / 2 \tag{B.11}$$

and the between-test-portion ranges as:

$$w_t = \left|x_{t,1} - x_{t,2}\right| \tag{B.12}$$

Calculate the general average:

$$\bar{\bar{x}} = \frac{1}{g}\sum_{t=1}^{g} \bar{x}_t \tag{B.13}$$

Estimate the standard deviation of sample averages:

$$s_{\bar{x}} = \sqrt{\sum_{t=1}^{g}\left(\bar{x}_t - \bar{\bar{x}}\right)^2 \Big/ (g-1)} \tag{B.14}$$

and the within-sample standard deviation:

$$s_w = \sqrt{\sum_{t=1}^{g} w_t^2 \Big/ (2g)} \tag{B.15}$$

where the summations in formulae B.13, B.14, and B.15 are over samples (t = 1, 2, ..., g).

Finally, estimate the between-sample standard deviation as:

$$s_s = \sqrt{\max\left(0, s_{\bar{x}}^2 - s_w^2 / 2\right)} \tag{B.16}$$

NOTE 1 The estimate of between-sample 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. In this case s_s = 0.

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

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

NOTE 3 An example is provided in Annex E.2

B.4 Procedures for checking stability

B.4.1 General considerations for checking stability

These clauses give guidance for meeting the stability requirements of section 6.1. The provisions of section 6.1.3 with regard to the properties to be studied apply to any experimental check on stability over the duration of the proficiency testing round and on stability during transport.

B.4.1.1 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 proficiency testing round, carried out during and after the round itself. In other circumstances, studies of transport effects and stability for the typical duration of a proficiency testing 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 check for evidence of instability by checking reported results for a trend with date of measurement.

B.4.1.2 The following considerations apply to 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.

— More than 2 proficiency test items should be tested if the variability between proficiency test items is large; more proficiency testing items or more replicates should be used if the repeatability is suspect (for example, if s_w or $s_r > 0{,}5\sigma_{pt}$).

NOTE ISO Guide 35 provides strategies for minimizing the effect on stability studies of long-term variation in the measurement process, such as isochronus studies or the use of stable reference materials.

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 round. 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, the comparison is between proficiency test items that are shipped with proficiency test items that are retained under controlled conditions.

NOTE 1 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.

NOTE 2 This model applies equally to proficiency testing schemes in testing and in calibration.

B.4.2.2 If a proficiency testing provider includes shipped proficiency test items in the stability assessment in B.4.2.1, 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.

B.4.2.3 A procedure for a basic stability check using measurements before and after a proficiency testing round is as follows:

a) Select a number $2g$ of the proficiency test items at random, where $g \geq 2$.

b) Select a single laboratory using a single measurement method with good intermediate precision.

c) Measure g proficiency test items before the planned date of distribution of proficiency test items to participants. Replicated measurements should be made in a fully randomised order.

d) Reserve the remaining g proficiency test items under conditions similar to the expected storage conditions at participants' premises.

e) As soon as reasonably possible after the closing date for return of participant results, measure the remaining g proficiency test items, using the same laboratory, measurement method and number of replicates as at a) above, with all replicates in a randomised order.

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

B.4.2.4 The following variations to the procedure in B.4.2.3 may be used:

a) The first group of g proficiency test items may be omitted if other measurements on the set of proficiency test items are available from the same laboratory and test method. For example, data from a prior homogeneity check may be used.

b) Conditions likely to accelerate change may be used to provide greater assurance of stability.

c) The second set of proficiency test items may additionally be subjected to conditions expected in shipping, in order to include a test of the effect of shipping.

d) Any other design and conditions that, together with the stability check criterion chosen, provides equal or greater assurance of stability may be used.

B.5 Assessment criterion for a stability check

B.5.1 Compare the general average of the measurements obtained in the check prior to distribution with the general average of the results obtained in the stability check. The proficiency test items may be considered to be adequately stable if:

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

B.5.2 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, then one of the following options should be taken:

a) use an isochronous stability study (see ISO Guide 35);

b) increase the uncertainty of the assigned value to account for possible instability;

c) expand the criterion for acceptance by adding the uncertainty of the difference to σ_{pt} using the following formula:

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

NOTE The factor of 2 in equation (B.18) is a coverage factor for the expanded uncertainty of the difference, providing approximately 95 % confidence, and the combined uncertainty calculation has intentionally assumed that $\bar{y}_1$ and $\bar{y}_2$ are independent.

B.5.3 If the criterion in equations (B.17) or (B.18) 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 proficiency test item preparation and storage procedures to see if improvements are possible; or

— do not evaluate participant performance.

B.5.4 The criterion at B.5.1 or B.5.2 may be replaced by an appropriate statistical test for a difference between the two sets of data provided that the test takes due account of replication and provides assurance of identifying stability at least equal to that provided by equation (B.18).

NOTE A t-test for significant difference at the 95% level of confidence, using the means for each proficiency test item, will usually provide similar or better assurance of detecting instability to equation (B.18) provided that the number of units tested is 3 or more.

B.6 Stability in transport conditions

B.6.1 The proficiency testing provider should check the effects of transport on proficiency testing items at least in the early stages of the proficiency testing scheme. Such a check should, where possible, compare proficiency test items retained at the proficiency testing provider's premises with proficiency test items subjected to shipping and return. Studies based on exposure to reasonably foreseeable conditions of transport, for example, may also be used.

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

B.6.3 Where the transport stability check involves the comparison of results for two groups of proficiency test items, one group being exposed to transport conditions and one group that is not, the criterion for sufficient stability in transport is the same as in section B.5.1 or B.5.2.

NOTE 1 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 will reflect any bias and increased variability (respectively) caused by transport conditions.

NOTE 2 An example of a stability check is shown in Annex E.2

Annex C
(normative)

Robust analysis

C.1 Robust analysis: Introduction

Interlaboratory comparisons present unique challenges for data analysis. While most interlaboratory comparisons provide unimodal and approximately symmetric data, most proficiency testing data sets include a proportion of results that are unexpectedly distant from the majority. These can arise for a variety of reasons; for example, from less experienced participants, from less precise, or perhaps new, measurement methods, or from participants who did not understand the instructions or who processed the proficiency test items incorrectly. Such outlying results can be highly variable and make conventional statistical techniques, including the mean and standard deviation, unreliable.

It is recommended (see 6.5.1) that proficiency testing providers use statistical techniques that are robust to outliers. Many such techniques have been proposed in the statistical literature, and many of those have been used successfully for proficiency testing. Most robust techniques additionally confer resistance to asymmetric outlier distributions.

This Annex describes several techniques that have been applied in proficiency testing and have different capabilities regarding robustness to contaminated populations (for example, efficiency and breakdown point), and differing simplicity of application. They are presented here in order of simplicity (simplest first, most complex last), which is approximately inversely related to efficiency because the more complex estimators tend to have been developed in order to improve efficiency.

NOTE 1 Annex D provides further information on efficiency, breakdown point and sensitivity to minor modes - three important indicators of the performance of various robust estimators.

NOTE 2 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.

C.2 Simple outlier-resistant estimators for the population mean and standard deviation

C.2.1 The median

The median is a simple and highly outlier-resistant estimator of the population mean for symmetric distributions. To determine the median, denoted $med(x)$:

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

$x\{1\}, x\{2\}, \ldots, x\{p\}$

ii) calculate

$$med\left(x\right) = \begin{cases} x_{\{(p+1)/2\}} & p \text{ odd} \\ \dfrac{\left[x_{\{p/2\}} + x_{\{1+p/2\}}\right]}{2} & p \text{ even} \end{cases} \quad \text{(C.1)}$$

C.2.2 Scaled median absolute deviation *MADe*

The scaled median absolute deviation *MADe*(x) provides an estimate of the population standard deviation for normally distributed data and is highly resistant to outliers. To calculate *MADe*(x):

i) Calculate the absolute differences d_i (for i = 1 to p) from

$$d_i = \left|x_i - med\left(x\right)\right| \quad \text{(C.2)}$$

ii) Calculate *MADe*(x) from

$$MADe(x) = 1{,}483\ med(d) \quad \text{(C.3)}$$

If 50 % or more of the participant results are the same, then *MADe*(x) will be zero, and it may be necessary to use the *nIQR* in section C.2.3, an arithmetic standard deviation (after outlier removal), or the procedure described in section C.5.2.

C.2.3 Normalized interquartile range *nIQR*

A robust estimator of the standard deviation similar to *MADe*(x) and slightly simpler to obtain 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. This statistic is commonly called the 'normalized InterQuartile Range' (or *nIQR*), and it is calculated as in formula (C.4):

$$nIQR(x) = 0{,}7413(Q_3(x) - Q_1(x)) \quad \text{(C.4)}$$

where

$Q_1(x)$ denotes the 25th percentile of x_i (i=1,2,...,p)

$Q_3(x)$ denotes the 75th percentile of x_i (i=1,2,...,p)

If the 75th and 25th percentiles are the same, the *nIQR* will be zero (as will *MADe*(x)) and an alternative procedure such as an arithmetic standard deviation (after outlier removal) or the procedure at C.5.2 should be used to calculate the robust standard deviation.

NOTE 1 The *nIQR* only requires sorting the data once compared to *MADe* but has breakdown point of 25 % (see Annex D), while *MADe* has breakdown point of 50 %. *MADe* can therefore tolerate an appreciably higher proportion of outliers than *nIQR*.

NOTE 2 Both *nIQR* and the *MADe* estimators show appreciable negative bias at p<30 which may adversely affect scores if these estimates are used in scoring participant results.

NOTE 3 Different statistical packages may use different algorithms for calculating quartiles, and therefore may produce slightly different *nIQR*.

NOTE 4 An example using simple robust estimators is included in Annex E.3.

C.3 Robust analysis: Algorithm A

C.3.1 Algorithm A with iterated scale

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

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

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

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

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

$$x^* = \text{median of } x_i \qquad (i = 1, 2, \ldots, p) \tag{C.5}$$

$$s^* = 1{,}483 \text{ median of } \left|x_i - x^*\right| \text{ with } (i = 1, 2, \ldots, p) \tag{C.6}$$

NOTE 1 Algorithms A and S given in this annex are reproduced from ISO 5725-5, with a slight addition to Algorithm A to specify a stopping criterion: no change in the 3rd significant figures of the robust mean and standard deviation.

NOTE 2 In some cases more than half of the results x_i will be 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:

$$\delta = 1{,}5s^* \tag{C.7}$$

For each x_i ($i = 1, 2, \ldots, p$), calculate:

$$x_i^* = \begin{cases} x^* - \delta & \text{when} \quad x_i < x^* - \delta \\ x^* + \delta & \text{when} \quad x_i > x^* + \delta \\ x_i & \text{otherwise} \end{cases} \tag{C.8}$$

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

$$x^* = \sum_{i=1}^{p} x_i^* / p \tag{C.9}$$

$$s^* = 1{,}134\sqrt{\sum_{i=1}^{p}\left(x_i^* - x^*\right)^2 / \left(p-1\right)} \tag{C.10}$$

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 equations C.7 to C.10, 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 mean and robust standard deviation (x^* and s^*). Alternative convergence criteria can be determined according to the design and reporting requirements for proficiency test results.

NOTE 3 Examples of use of Algorithm A with iterated scale are provided in Annex E.1 and E.3.

C.3.2 Variants of Algorithm A

Algorithm A with iterated scale in section C.3.1 has modest breakdown (approximately 25 % for large data sets[25]) and the starting point for s^* suggested in C.3.1 for data sets where $MADe(x)$ is zero can seriously degrade outlier resistance when there are severe outliers in the data set. The following variations should be considered where the proportion of outliers is expected to be over 20 % in any data set or where the initial value for s^* is adversely affected by extreme outliers:

i) Replace $MADe$ with $med\left(\left|x_i - \bar{x}\right|\right)$ when $MADe$=0, or use an alternative estimator such as that described in C.5.1 or the arithmetic standard deviation (after outlier removal).

ii) Where the robust standard deviation is not used in scoring, use $MADe$ (amended as i) above) and do not update s^* during iteration. Where the robust standard deviation is used in scoring, replace s^* with the Q estimator described in C.5 and do not update s^* during iteration.

NOTE Variant ii) improves the breakdown point of Algorithm A to 50 % [25], allowing the algorithm to cope with a higher proportion of outliers.

C.4 Robust analysis: 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 standard deviations or ranges, sorted into increasing order, by:

$$w_{\{1\}}, w_{\{2\}}, \ldots, w_{\{p\}}$$

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

Calculate an initial value for w^* as:

$$w^* = \text{median of } w_i \quad (i = 1, 2, \ldots, p) \tag{C.11}$$

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, 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:

$$\psi = \eta \times w^* \tag{C.12}$$

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

$$w_i^* = \begin{cases} \psi & \text{if } w_i > \psi \\ w_i & \text{otherwise} \end{cases} \tag{C.13}$$

Calculate the new value of w^* from:

$$w^* = \xi\sqrt{\sum_{i=1}^{p}\left(w_i^*\right)^2 / p} \tag{C.14}$$

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.

NOTE Algorithm S provides an estimate of the population standard deviation when supplied with standard deviations from a single normal distribution (and hence provides an estimate of the repeatability standard deviation when the assumptions of ISO 5725-2 apply).

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 intensive robust estimators: *Q* method and Hampel estimator

C.5.1 Rationale for computationally intensive estimators

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 become unreliable when more than 20 % of results are outliers, or where there are bimodal (or multimodal) distributions, and some may become unacceptably variable for smaller numbers of participants. Further, none can handle replicated data from participants. 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, and the robust mean and/or standard deviation are used for scoring there is a considerable risk for misclassifying participants due to the use of ineffective statistical methods.

Robust techniques that combine good efficiency (that is, comparatively low variability) with tolerance for a high proportion of outliers tend to be more complex and require more computational resources, but the techniques are referenced in available literature and International Standards. Some of these additionally provide useful performance gains when the underlying distribution of data is skewed or when some results are quoted as below a detection or reporting limit.

The following paragraphs describe some high-efficiency, high-breakdown methods for estimating standard deviation and location (mean) that are useful for data with larger proportions of outliers and that show lower variability than simpler estimators. One of the estimators described can also be used to estimate a reproducibility standard deviation when participants report multiple observations.

C.5.2 Determination of a robust standard deviation using Q and Q_n methods

C.5.2.1 Q_n [34] is a high-breakdown, high-efficiency estimator of the population standard deviation which is unbiased for normally distributed data (that is, under the assumption that there are no outliers). Q_n uses a single reported result (including a mean or median of replicates) for each participant. The calculation relies on the use of pairwise differences within the data set and therefore it is not dependent on an estimate of the mean or median of the data. The implementation described here includes corrections to ensure that the estimate is unbiased for all practical data set sizes.

To calculate Q_n for a data set ($x_1, x_2, \ldots x_p$) with p reported results:

i) Calculate the $p(p\text{-}1)/2$ absolute differences

$$d_{ij} = \left| x_i - x_j \right| \text{ for } i = 1,2\ldots p-1 \text{ and } j = i+1, i+2\ldots p \tag{C.15}$$

ii) Denote the ordered differences d_{ij} by

$$d_{\{1\}}, d_{\{2\}} \ldots d_{\{p(p\text{-}1)/2\}} \tag{C.16}$$

iii) Calculate

$$k = \frac{h(h-1)}{2} \tag{C.17}$$

that is, k is the number of distinct pairs chosen from h objects, where:

$$h = \begin{cases} p/2 & p \text{ even} \\ (p-1)/2 & p \text{ odd} \end{cases} \tag{C.18}$$

iv) Calculate Q_n as

$$Q_n = 2,2219 d_{\{k\}} b_p \tag{C.19}$$

where b_p is selected from Table C.2 for a particular number p of data points or, for $p > 12$, is calculated from

$$b_p = \frac{1}{r_p + 1} \tag{C.20}$$

where

$$r_p = \begin{cases} \frac{1}{p}\left[1,6019 + \frac{1}{p}\left(-2,128 - \frac{5,172}{p}\right)\right] & p \text{ odd} \\ \frac{1}{p}\left[3,6756 + \frac{1}{p}\left(1,965 + \frac{1}{p}\left(6,987 - \frac{77}{p}\right)\right)\right] & p \text{ even} \end{cases} \quad \text{(C.21)}$$

NOTE 1 The factor of 2,2219 is a correction factor to give an unbiased estimate of standard deviation for large p. The correction factors b_p for small p are in table C.2 and the calculation for r_p for $p > 12$ are as provided in reference [34] from extensive simulation and subsequent regression analysis.

NOTE 2 The simple algorithm described above requires considerable computing resources for larger data sets, for example $p > 1000$. A fast and memory-efficient implementation capable of handling much larger data sets has been published with full computer code [34] for use with larger data sets; reference [34] cited acceptable performance for p over 8000 at the time of publication.

Table C.2 — Correction factor b_p for $2 \leq p \leq 12$

p	2	3	4	5	6	7	8	9	10	11	12
b_p	0,9937	0,9937	0,5132	0,8440	0,6122	0,8588	0,6699	0,8734	0,7201	0,8891	0,7574

C.5.2.2 The Q method produces a high-breakdown, high-efficiency estimate of the standard deviation of proficiency testing results reported by different laboratories. The Q method is not only robust against outlying results, but also against a situation where many test results are equal, e.g. due to quantitative data on a discontinuous scale or due to rounding distortions. In such a situation other Q-like methods can fail because many pairwise differences are zero.

The Q method can be used for proficiency testing both with single results per participant (including a mean or median of replicates) and for replicates. The direct use of replicates in the calculation improves the efficiency of the method.

The calculation relies on the use of pairwise differences within the data set and is therefore not dependent on an estimate of the mean or median of the data. The method is known as Q/Hampel when it is used together with the finite step algorithm for the Hampel estimator described in C.5.3.3.

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

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

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

$$H_1(x) = \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{m=1}^{n_j} \mathbf{I}\left\{\left|y_{ik} - y_{jm}\right| \leq x\right\} \quad \text{(C.22)}$$

where $\mathbf{I}\left\{\left|y_{ik} - y_{jm}\right| \leq x\right\} = \begin{cases} 1 & \text{if } \left|y_{ik} - y_{jm}\right| \leq x \\ 0 & \text{otherwise} \end{cases}$ denotes the indicator function.

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

$x_1, \ldots, x_r$, where $x_1 < x_2 < \ldots < x_r$.

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

$$G_1\left(x_i\right) = \begin{cases} 0,5 \cdot \left(H_1\left(x_i\right) + H_1\left(x_{i-1}\right)\right) & \text{if} \quad i \geq 2 \\ 0,5 \cdot H_1\left(x_1\right) & \text{if} \quad i = 1; x_1 > 0 \end{cases} \tag{C.23}$$

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 $0 \leq x_1 < x_2 < \ldots < x_r$.

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

$$s^* = \frac{G_1^{-1}\left(0,25 + 0,75 \cdot H_1\left(0\right)\right)}{\sqrt{2}\Phi^{-1}\left(0,625 + 0,375 \cdot H_1\left(0\right)\right)} \tag{C.24}$$

where $H_1(0)$ is calculated as in equation (C.22) and is equal to zero unless there are exact ties in the data set, and where $\Phi^{-1}(q)$ is the q^{th} quantile of the standard normal distribution.

NOTE 1 This algorithm does not depend on a mean value; it can be used together with either a value from combined participant results or a specified reference value.

NOTE 2 Other variants of the Q method provide robust estimates of both repeatability and reproducibility standard deviation [25,34].

NOTE 3 The theoretical basis for the Q method, including asymptotic performance and finite sample breakdown, are described in references [26] and [34].

NOTE 4 If the underlying data of the participants represent single measurement results obtained with one specific measurement method, the robust standard deviation is an estimate of the reproducibility standard deviation as in equation (C.21).

NOTE 5 The reproducibility standard deviation is not necessarily the most appropriate standard deviation for use in proficiency testing because it is usually an estimate of the dispersion of single results and not an estimate of the dispersion of means or medians of replicated results from each participant. However the dispersion of means or medians of replicated results is only slightly below the dispersion of single results of different laboratories, if the ratio of reproducibility standard deviation divided by the repeatability standard deviation is greater than 2. If this ratio is below 2, for scoring in proficiency testing it may be considered to replace the reproducibility standard deviation s_R by the corrected value $\sqrt{s_R^2 - \frac{m-1}{m} s_r^2}$, where m denotes number of replicates and s_r^2 the repeatability variance as calculated in [35], or to use not the replicates but the mean of replicates per participant for the Q method.

NOTE 6 Note 5 applies only if the scoring is conducted on the basis of means or medians of replicated results. If the replicates are blind replicate proficiency test items, scores should be given for each replicate. In this case the reproducibility standard deviation is the most appropriate standard deviation.

NOTE 7 An example to which the Q method has been applied is shown in Annex E.3.

C.5.3 Determination of a robust mean using the Hampel estimator

C.5.3.1 The Hampel estimate is a highly robust and efficient estimate of the overall mean of results reported by different laboratories. As there is no explicit formula for obtaining the Hampel estimate, in this paragraph two algorithms are provided. The first one can be more easily implemented but may lead

to deviating results in different implementations. The second one provides unique results depending only on the underlying standard deviation.

C.5.3.2 The following calculation provides an iterative reweighting scheme for obtaining the Hampel estimate of location.

i) Denote the data as $x_1, x_2 \ldots x_p$

ii) Set x^* to $med(x)$ (section C.2.1)

iii) Set s^* to a suitable robust estimate of standard deviation, for example $MADe$, Q_n or s^* from the Q method.

iv) For each data point x_i, calculate q_i from

$$q_i = \left| \frac{x_i - x^*}{s^*} \right|$$

v) Calculate weights w_i from

$$w_i = \begin{cases} 0 & |q| > 4,5 \\ (4,5 - q)/q & 3 < |q| \leq 4,5 \\ 1,5 / q & 1,5 < |q| \leq 3,0 \\ 1 & |q| \leq 1,5 \end{cases}$$

vi) Recalculate x^* from

$$x^* = \frac{\sum_{i=1}^{p} w_i x_i}{\sum_{i=1}^{p} w_i}$$

vii) Repeat steps iv) to vi) until x^* converges. Convergence may be assumed when the change in x^* from one iteration to the next is less than $0,01\,s^*/\sqrt{p}$, corresponding to approximately 1 % of the standard error in x^*. Other, more precise, convergence criteria may be used.

This implementation of the Hampel estimator is not guaranteed to have a unique solution or to result in the best solution because a poor choice of initial location x^* and/or s^* may exclude important parts of the data set. The proficiency testing provider should accordingly implement measures to check for the possibility of a poor solution or provide unambiguous rules for choice of location. The most common rule is to choose the solution nearest the median. Reviewing the results to ensure that no large proportion of the data set is outside the range $|q|>4.5$ can also assist in confirming a viable solution.

NOTE 1 This implementation of Hampel's estimator has approximately 96 % efficiency for normally distributed data.

NOTE 2 An example using this implementation is given in Annex E.3

NOTE 3 Hampel's estimator may be tuned for greater efficiency or greater resistance to outliers by changing the weight function. The general form of the weighting function is

$$w_i = \begin{cases} 0 & |q| > c \\ a(c-q)/[q(c-b)] & b < |q| \le c \\ a/q & a < |q| \le b \\ 1 & |q| \le a \end{cases}$$

where a, b and c are tuning parameters. For the implementation here, a = 1,5, b = 3,0 and c = 4,5. Greater efficiency is obtained by increasing the range; improved resistance to outliers or minor modes is obtained by reducing the range.

C.5.3.3 The following finite step algorithm yields the Hampel estimate of location without iterative reweighting[25].

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

$$\sum_{i=1}^{p} \Psi\left(\frac{y_i - x^*}{s^*}\right) = 0 \tag{C.25}$$

where

$$\Psi(q) = \begin{cases} 0 & q \le -4,5 \\ -4,5 - q & -4,5 < q \le -3 \\ -1,5 & -3 < q \le -1,5 \\ q & -1,5 < q \le 1,5 \\ 1,5 & 1,5 < q \le 3 \\ 4,5 - q & 3 < q \le 4,5 \\ 0 & q > 4,5 \end{cases} \tag{C.26}$$

and s^* is the robust 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.25) (interpreted here as a function of x^*) are as follows:

Calculate all interpolation nodes

— for the first value y_1:

$$d_1 = y_1 - 4,5 \cdot s^*,\ d_2 = y_1 - 3 \cdot s^*,\ d_3 = y_1 - 1,5 \cdot s^*,\ d_4 = y_1 + 1,5 \cdot s^*,\ d_5 = y_1 + 3 \cdot s^*,\ d_6 = y_1 + 4,5 \cdot s^*$$

— for the second value y_2:

$$d_7 = y_2 - 4,5 \cdot s^*,\ d_8 = y_2 - 3 \cdot s^*,\ d_9 = y_2 - 1,5 \cdot s^*,\ d_{10} = y_2 + 1,5 \cdot s^*,\ d_{11} = y_2 + 3 \cdot s^*,\ d_{12} = y_2 + 4,5 \cdot s^*$$

— and so on for all values $y_3, \ldots, y_p$.

Sort these data $d_1, d_2, d_3, \ldots, d_{6 \cdot p}$ in ascending order, $d_{\{1\}}, d_{\{2\}}, d_{\{3\}}, \ldots, d_{\{6 \cdot p\}}$

Then calculate for each $m = 1, \ldots, (6 \cdot p - 1)$

$$p_m = \sum_{i=1}^{p} \Psi\left(\frac{y_i - d_{\{m\}}}{s^*}\right)$$

and check whether

(i) $p_m = 0$. If so, $d_{\{m\}}$ is a solution of equation (C.25).

(ii) $p_{m+1} = 0$. If so, $d_{\{m+1\}}$ is a solution of equation (C.25).

(iii) $p_m \cdot p_{m+1} < 0$. If so, $x_m = d_{\{m\}} - \dfrac{p_m}{\dfrac{p_{m+1} - p_m}{d_{(m+1)} - d_{(m)}}}$ is a solution of equation (C.25).

Let S denote the set of all of these solutions of equation (C.25).

The solution $x^* \in S$ nearest the median is used as location parameter x^*, i.e.

$$\left|x^* - med\left(y_1, y_2, \ldots, y_p\right)\right| = \min\left\{\left|x - med\left(y_1, y_2, \ldots, y_p\right)\right|; x \in S\right\}$$

Several solutions may exist. If there are two solutions nearest the median, or if there is no solution at all, the median itself is used as location parameter x^*.

NOTE 1 This implementation of Hampel's estimator has approximately 96 % efficiency for normally distributed data.

NOTE 2 If this estimation method is used, laboratory results differing from the mean by more than 4,5 times the reproducibility standard deviation no longer have any effect on the calculation result, i.e. they are treated as outliers.

C.5.4 The Q/Hampel method

The method known as Q/Hampel uses the Q method described in C.5.3.2 for the calculation of the robust standard deviation s^* together with the finite step algorithm for the Hampel estimator described in C.5.3.3 for the calculation of the location parameter x^*.

When participants report multiple observations, the Q method described in C.5.3.2 is used for the calculation of the robust reproducibility standard deviation s_R. For the calculation of the robust repeatability standard deviation s_r a second algorithm using the pairwise differences within the laboratories is applied.

NOTE A web application for the Q/Hampel method is available[37].

C.6 Other robust techniques

The methods described in this Annex do not constitute a comprehensive collection of valid approaches, and none is guaranteed to be optimal for all situations. Other robust estimators may be used at the discretion of the proficiency testing provider, subject to demonstration, by reference to known efficiency, breakdown point and any other appropriate properties, that they fulfil the particular requirements of the proficiency testing scheme.

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, performance is evaluated using the pre-determined assigned value and performance criterion, so proficiency testing can be conducted with just one participant. This type of interlaboratory comparison can be called a bilateral comparison, or measurement audit, and can be very useful in many situations, for example, in calibration.

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 determined using participant results.

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 the Grubbs test for a single outlier and for two simultaneous outliers in the same direction. Grubbs and other 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 dispersion perform acceptably for p≥ 12.

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 situation using a simple mean as the assigned value and a standard deviation of results as the standard deviation for proficiency assessment, this criterion cannot be met for a normal distribution with p≤12, after any removal of outliers. For use

of the median as the assigned value (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 7.7.3 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 dispersion for robust location estimators), or where the proficiency testing 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 standard deviation described in section C.5 is considerably more efficient than either the *MADe* or *nIQR* from Annex C.1.

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

— $p=2$: use $|x_1-x_2|/\sqrt{2}$;

— $p=3$, locations and scale unknown: use *MADe* to protect against excessively high estimates of the standard deviation or the mean absolute deviation to protect against unduly small estimates of the standard deviation, for example when rounding may give two identical values;

— $p \geq 4$: A specific *M*-estimate of standard deviation based on a logarithmic weighting function was recommended by reference [27]; 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

$$s^* = \frac{1}{0,798 \times p} \sum_{i=1}^{p} \left| x_i - med(x) \right| \qquad \text{(D.1)}$$

D.2 Efficiency and breakdown points for robust procedures

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

Breakdown point — the proportion of values in the data set that can be replaced by arbitrarily large values without the estimate also becoming arbitrarily large.

Efficiency — the ratio of the estimator variance divided by the variance of a minimum variance estimator for the distribution in question.

Resistance to minor modes — the ability of an estimator to resist the bias caused by a minority group of discrepant results (typically less than 20 % of the data set).

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 contaminating 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

The breakdown point is the proportion of values in the data set that can be outliers without the estimate being adversely affected. The breakdown point is a measure of resistance to outlying values; high breakdown point is associated with resistance to a high proportion of outliers. Breakdown points and resistance to minor modes for the estimators in Annex C are presented in Table D.1. 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	Resistance to Minor Modes
Sample mean	Mean	0 %	Poor
Sample standard deviation	Standard deviation	0 %	Poor
Sample median	Mean	50 %	Good
nIQR	Standard deviation	25 %	Moderate
MADe	Standard deviation	50 %	Moderate - Good
Algorithm A	Mean and Standard deviation	25 %	Moderate
Q_n and Q / Hampel	Mean and Standard deviation	50 %	Moderate (Very Good for minor modes more distant than 6 s*)

NOTE The definition of breakdown point used here is the proportion of a large normally distributed data set that can be moved to +infinity without the estimate also moving to infinity. For example, if just under 50 % of a data set is replaced by +infinity, the median will remain within the remaining finite data.

In summary, the sample mean and standard deviation can break down with only a single outlier. The robust methods using the median, *MADe*, and Q/Hampel methods can tolerate a very large proportion of outliers. Algorithm A with iterated standard deviation and *nIQR* have a breakdown point of 25 %. In any situation with a large proportion of outliers (>20 %), any conventional or robust procedure can produce unreasonable estimates of location and dispersion, and caution should be used in interpretation of such values.

D.2.3 Relative efficiency

All estimates have sampling variance – that is, the estimates can vary from round to round of a proficiency testing scheme, 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 sample 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 round to round, even if there are no outliers or different subgroups of participants. Table D.2 provides relative efficiencies for the estimators presented in Annex C.

Table D.2 — Relative efficiency of robust estimators for the population mean and standard deviation, for normally distributed datasets with n=50 or 500 participants:

Statistical Estimator	Mean, n=50	Mean, n=500	SD, n=50	SD, n=500
Sample mean and Standard deviation	100 %	100 %	100 %	100 %
Median and *nIQR*	66 %	65 %	38 %	37 %
Median and *MADe*	66 %	65 %	37 %	37 %
Algorithm A	97 %	97 %	74 %	73 %
Q_n and Q / Hampel	96 %	96 %	73 %	81 %

These results demonstrate that there is no statistical method that is perfect for all situations. The sample mean and standard deviation are optimal with a normal distribution but break down in case of outliers. Simple robust methods such as median, *MADe* or *nIQR* perform comparatively poorly for normally distributed data but can be effective when outliers are present or the data set is small.

D.3 Use of proficiency testing data for evaluating the reproducibility and repeatability of a measurement method

D.3.1 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 is possible to use the results of proficiency testing schemes to verify, and perhaps establish the repeatability and reproducibility of a measurement method [15] when the proficiency testing scheme meets the following conditions:

a) the proficiency testing items 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 proficiency testing round, and their competence is not placed in doubt by the results in the round.

D.3.2 In order to provide sufficient data for evaluation of repeatability and reproducibility of a test method from a proficiency testing scheme, the following design conditions shall be used:

a) a sufficient number of participants to satisfy a collaborative study have demonstrated competence with a measurement method on previous rounds of a proficiency testing scheme, and have committed to follow the measurement method without modification;

b) where repeatability is to be assessed, each proficiency testing round used for the repeatability assessment should include at least two proficiency test items or a requirement for replicate observations;

c) where practicable, participants should be provided with separately identified blind replicates rather than being requested to perform replicate measurements on the same proficiency test item;

d) proficiency test items used in one or several rounds of the proficiency testing scheme cover the range of levels and types of routine samples for which the measurement method is intended;

e) data analysis procedures used for assessing repeatability and reproducibility should be consistent with ISO 5725 or the collaborative study protocol in use.

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. Specific examples should not be considered to be recommendations for use in particular proficiency testing schemes.

E.1 Effect of censored values (section 5.5.3.3)

Table E.1 shows 23 results for a round of a proficiency testing scheme, of which 5 results are indicated as 'Less Than' some amount. The robust mean (x^*) and standard deviation (s^*) from Algorithm A 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 are indicated with '#'. This assumes that the evaluation would be 'unacceptable' (action signal) for any result where the quantitative part is outside the $x^* \pm 3s^*$. 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 x '<' value
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

Table E.1 *(continued)*

Participant	Result	'<' ignored	'< ' deleted	0,5 x '<' value
Summary				
Number of Results	23	23	18	23
x*		26,01	26,81	23,95
s*		7,23	5,29	8,60

The choice of how to handle the "less than" samples has a significant effect on the robust mean and standard deviation, and on the performance evaluation. The proficiency testing provider is expected to determine an appropriate method.

E.2 Homogeneity and Stability test – Arsenic (As) in chocolate (section 6.1)

Proficiency test items are prepared for use in an international proficiency test, and then for subsequent use as reference materials. 1000 vials are manufactured.

Homogeneity check: 10 proficiency test items are selected using a stratified random selection of proficiency test items 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.2 below. The procedure in Annex B.3 is followed, resulting in the summary statistics listed. The fitness-for purpose σ_{pt} for As in chocolate is 15 %, so the estimate of sample variability is checked against 0,3 times the σ_{pt}.

Table E.2 — Homogeneity data for proficiency test items of arsenic in chocolate

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	0,180
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 x 0,15 = 0,02807
Check value:	$0{,}3\sigma_{pt}$ = 0,00842
s_s:	is less than the check value, so homogeneity is sufficient.

Stability check: 2 proficiency test items are randomly selected and stored at an elevated temperature (60 °C) for the duration of the round of the proficiency testing scheme (6 weeks). The proficiency test items were tested in duplicate (Table E.3), and the four results are checked against the homogeneity values.

Table E.3 — Stability data for proficiency test items for arsenic in chocolate

Stability sample	Replicate 1	Replicate 2
164	0,191	0,198
732	0,190	0,196

Overall average:	= 0,19375
Difference from Homogeneity mean:	0,19375 - 0,18715 = 0,00660
Check value:	$0{,}3\sigma_{pt}$ = 0,00842

Difference is less than the check value, so stability is sufficient.

E.3 Comprehensive Example of Atrazine in Drinking Water

A proficiency testing scheme for a herbicide (Atrazine) in drinking water has 34 participants. This raw data as submitted in Table E.4, ordered by value for clarity. The Table shows calculated values for the robust mean and standard deviation following Algorithm A, following 6 iterations until the robust mean and standard deviation do not change at their third significant figures. The data are shown as ranked data plot in Figure E.1 and in corresponding histogram and kernel density plot in Figures E.2 and Figure E.3, respectively.

Table E.5 shows the estimates of location (average) and standard deviation using various classical and robust techniques. The uncertainty of the estimate of location is also shown. The statistics for bootstrap method are derived from the procedures in references [17,18] and the R software package [see R3.1.1 below]. Figure E.4 shows the different estimates of location and the estimate of expanded uncertainty ($2u(x_{pt})$) as the error bar.

Table E.4 — Calculation of the robust average and standard deviation for Atrazine in drinking water

	x_i	1st iteration	2nd iteration	3rd iteration	4th iteration	5th iteration	6th iteration
x* - δ		**0,204163**	**0,199732**	**0,198466**	**0,198037**	**0,197865**	**0,197790**
x* + δ		**0,319837**	**0,315969**	**0,315871**	**0,316065**	**0,316185**	**0,316243**
1	0,0400	0,2042	0,1997	0,1985	0,1980	0,1979	0,1978
2	0,0550	0,2042	0,1997	0,1985	0,1980	0,1979	0,1978
3	0,1780	0,2042	0,1997	0,1985	0,1980	0,1979	0,1978
4	0,2020	0,2042	0,2020	0,2020	0,2020	0,2020	0,2020
5	0,2060	0,2060	0,2060	0,2060	0,2060	0,2060	0,2060
6	0,2270	0,2270	0,2270	0,2270	0,2270	0,2270	0,2270
7	0,2280	0,2280	0,2280	0,2280	0,2280	0,2280	0,2280
8	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300
9	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300
10	0,2350	0,2350	0,2350	0,2350	0,2350	0,2350	0,2350
11	0,2360	0,2360	0,2360	0,2360	0,2360	0,2360	0,2360
12	0,2370	0,2370	0,2370	0,2370	0,2370	0,2370	0,2370
13	0,2430	0,2430	0,2430	0,2430	0,2430	0,2430	0,2430
14	0,2440	0,2440	0,2440	0,2440	0,2440	0,2440	0,2440
15	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450
16	0,2555	0,2555	0,2555	0,2555	0,2555	0,2555	0,2555
17	0,2600	0,2600	0,2600	0,2600	0,2600	0,2600	0,2600

Table E.4 *(continued)*

	x_i	1st iteration	2nd iteration	3rd iteration	4th iteration	5th iteration	6th iteration
x* - δ		**0,204163**	**0,199732**	**0,198466**	**0,198037**	**0,197865**	**0,197790**
x* + δ		**0,319837**	**0,315969**	**0,315871**	**0,316065**	**0,316185**	**0,316243**
18	0,2640	0,2640	0,2640	0,2640	0,2640	0,2640	0,2640
19	0,2670	0,2670	0,2670	0,2670	0,2670	0,2670	0,2670
20	0,2700	0,2700	0,2700	0,2700	0,2700	0,2700	0,2700
21	0,2730	0,2730	0,2730	0,2730	0,2730	0,2730	0,2730
22	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740
23	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740
24	0,2780	0,2780	0,2780	0,2780	0,2780	0,2780	0,2780
25	0,2811	0,2811	0,2811	0,2811	0,2811	0,2811	0,2811
26	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870
27	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870
28	0,2880	0,2880	0,2880	0,2880	0,2880	0,2880	0,2880
29	0,2890	0,2890	0,2890	0,2890	0,2890	0,2890	0,2890
30	0,2950	0,2950	0,2950	0,2950	0,2950	0,2950	0,2950
31	0,2960	0,2960	0,2960	0,2960	0,2960	0,2960	0,2960
32	0,3110	0,3110	0,3110	0,3110	0,3110	0,3110	0,3110
33	0,3310	0,3198	0,3160	0,3159	0,3161	0,3162	0,3162
34	0,4246	0,3198	0,3160	0,3159	0,3161	0,3162	0,3162
average	0,2512	0,2579	0,2572	0,2571	0,2570	0,2570	0,2570
SD	0,0672	0,0342	0,0345	0,0347	0,0348	0,0348	0,.0348
δ		0,0578	0,0581	0,0587	0,0590	0,0592	0,0592
New x*	0,2620	0,2579	0,2572	0,2571	0,2570	0,2570	**0,2570**
New s*	0,0386	0,0387	0,0391	0,0393	0,0394	0,0395	**0,0395**

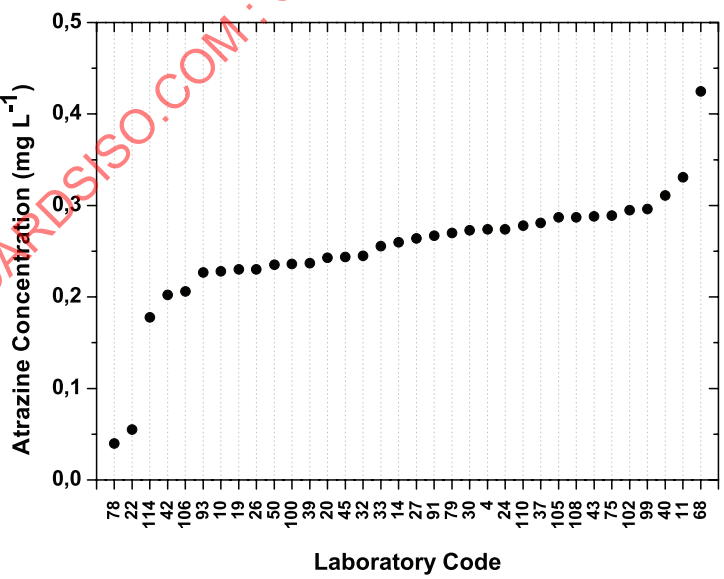

Figure E.1 — – Ranked participant results for Atrazine (data from Table E.4)