



BSI Standards Publication

Statistical methods for use in proficiency testing by interlaboratory comparison

National foreword

This British Standard is the UK implementation of [ISO 13528](#). It supersedes [BS ISO 13528:2015](#), which is withdrawn.

The UK participation in its preparation was entrusted to Technical Committee SS/6, Statistical procedures for measurement methods and results.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

© The British Standards Institution 2022
Published by BSI Standards Limited 2022

ISBN 978 0 539 05595 5

ICS 03.120.30

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 August 2022.

Amendments/corrigenda issued since publication

Date	Text affected
------	---------------

INTERNATIONAL STANDARD

ISO
13528

Third edition
2022-08-18

Statistical methods for use in proficiency testing by interlaboratory comparison

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



Reference number
ISO 13528:2022(E)
Documents
Document Sharing Hub
© ISO 2022



COPYRIGHT PROTECTED DOCUMENT

© ISO 2022, Published in Switzerland

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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	v
0 Introduction	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 General principles	4
4.1 General requirements for statistical methods	4
4.2 Basic model	5
4.3 General approaches for the evaluation of performance	5
5 Guidelines for the statistical design of proficiency testing schemes	6
5.1 Introduction to the statistical design of proficiency testing schemes	6
5.2 Basis of a statistical design	6
5.3 Considerations for the statistical distribution of results	7
5.4 Considerations for small numbers of participants	8
5.5 Guidelines for choosing the reporting format	8
5.5.1 General requirements for reporting format	8
5.5.2 Reporting of replicate measurements	9
5.5.3 Reporting of 'less than' or 'greater than' a limit (censored data)	9
5.5.4 Number of significant digits	9
6 Guidelines for the initial review of proficiency testing items and results	10
6.1 Homogeneity and stability of proficiency test items	10
6.2 Considerations for different measurement methods	11
6.3 Blunder removal	11
6.4 Visual review of data	12
6.5 Robust statistical methods	12
6.6 Outlier techniques for individual results	13
7 Determination of the assigned value and its standard uncertainty	14
7.1 Choice of method of determining the assigned value	14
7.2 Determining the uncertainty of the assigned value	14
7.3 Formulation	15
7.4 Certified reference material	16
7.5 Results from one laboratory	16
7.6 Consensus value from expert laboratories	17
7.7 Consensus value from participant results	18
7.8 Comparison of the assigned value with an independent reference value	19
8 Determination of criteria for evaluation of performance	20
8.1 Approaches for determining evaluation criteria	20
8.2 By perception of experts	21
8.3 By experience from previous rounds of a proficiency testing scheme	21
8.4 By use of a general model	21
8.5 Using the repeatability and reproducibility standard deviations from a previous collaborative study of precision of a measurement method	22
8.6 From data obtained in the same round of a proficiency testing scheme	22
8.7 Monitoring interlaboratory agreement	23
9 Calculation of performance statistics	24
9.1 General considerations for determining performance	24
9.2 Limiting the uncertainty of the assigned value	24
9.3 Estimates of deviation (measurement error)	25
9.4 z scores	26
9.5 z' scores	27
9.6 Zeta scores (ζ)	28

9.7	E_n scores	29
9.8	Evaluation of participant uncertainties in testing	30
9.9	Combined performance scores	31
10	Graphical methods for describing performance scores	31
10.1	Application of graphical methods	31
10.2	Histograms of results or performance scores	32
10.3	Kernel density plots	32
10.4	Bar-plots of standardized performance scores	34
10.5	Youden plot	34
10.6	Plots of repeatability standard deviations	35
10.7	Split samples	36
10.8	Graphical methods for combining performance scores over several rounds of a proficiency testing scheme	36
11	Design and analysis of qualitative proficiency testing schemes (including nominal and ordinal properties)	37
11.1	Types of qualitative data	37
11.2	Statistical design	38
11.3	Assigned values for qualitative proficiency testing schemes	38
11.4	Performance evaluation and scoring for qualitative proficiency testing schemes	40
Annex A (normative) Symbols		42
Annex B (informative) Homogeneity and stability of proficiency test items		44
Annex C (informative) Robust analysis		52
Annex D (informative) Additional guidance on statistical procedures		63
Annex E (informative) Illustrative examples		68
Annex F (Informative) Example of computer code for plotting and resampling analysis ("bootstrapping") of PT results		91
Bibliography		92

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 third edition of [ISO 13528](#) cancels and replaces the second edition ([ISO 13528:2015](#)), of which it constitutes a minor revision. The changes are as follows:

- notes have been added to [10.1](#), [10.4.3](#) and [10.5.3](#) to draw attention to additional graphical techniques that can assist in meeting the provisions of [10.1](#);
- [Formulae B.4](#) and [B.8](#) have been corrected to use s_t^2 instead of w_t^2 ;
- [Formula B.16](#) has been corrected so that the term inside the square root is always non-negative;
- in [Table C.2](#), the correction factor associated with $p = 2$ has been corrected to read 0,3994;
- additional literature references to the source of values in [Table C.2](#) have been added to the Bibliography and referenced from Notes 1 and 2 of [C.5.2.1](#);
- font styles (Italic or Roman) have been amended throughout for consistency in formulae.

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](#). 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 shall 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 document, both strategies are provided for. However, attention is drawn to the discussion that will be found in [Clauses 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](#)

This document 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. [ISO/IEC 17043:2010](#), Annex B, briefly describes the general statistical methods that are used in proficiency testing schemes. This document 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 this document, with the notes that describe different types of proficiency testing and the range of designs that can be used. This document cannot specifically cover all purposes, designs, matrices and measurands. The techniques presented in this document 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.

This document 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. The revision of this document 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, including [ISO/IEC 17043:2010](#), Annex B, which includes considerations for qualitative results.

This document 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](#) 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 document can specify further what that necessary expertise is. For some applications an advanced degree in statistics is useful, but usually the needs for expertise can be met by individuals with technical expertise in other areas, who are familiar with basic statistical concepts and have experience or training in the common techniques applicable to the analysis of data from proficiency testing schemes. If an individual is responsible for 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 document 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 document can be accomplished by conventional spread sheet applications, perhaps with customised routines for a particular proficiency testing scheme or analysis; some techniques will require computer applications that are freely available. In all cases, the users are expected to verify the validity and accuracy of their calculations, especially when special routines have been entered by the user. However, even when the techniques in this document 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 understand the nature of the applications and the statistical assumptions, and to identify and investigate anomalies that can occur in any round of a proficiency testing scheme.

This page deliberately left blank

Statistical methods for use in proficiency testing by interlaboratory comparison

1 Scope

This document 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 document provides recommendations on the interpretation of proficiency testing data by participants in such proficiency testing schemes and by accreditation bodies.

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

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

NOTE The procedures in this document can also be applied for the assessment of expert opinion where the opinions or judgments are reported in a form which can 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 document that relate to nominal (qualitative) properties can be applied.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

[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 Guide 30](#), *Reference materials — Selected terms and definitions*

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 3534-1 ISO 3534-2 apply. Mathematical symbols are listed in [Annex A](#).

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <https://www.electropedia.org/>

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* (3.1)

Note 1 to entry: For the purposes of this document, 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* (3.2)

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* (3.5), 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* (3.3) and the *standard deviation for proficiency assessment* (3.4)

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* (3.3) and the combined standard uncertainties for the result and the *assigned value* (3.3)

3.9

proportion of allowed limit score

standardized measure of performance, calculated using the participant result, *assigned value* (3.3) and the criterion for *measurement error* (3.5) 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 of 3 or greater 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* (3.1)

Note 1 to entry: The phrase 'consensus value' is typically used to describe estimates of location and dispersion derived from participant results in a round of a proficiency testing scheme, 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 proficiency testing 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* (3.2) 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* (3.13) performance in *proficiency testing* (3.2)

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* (3.2) scheme

3.16

proficiency testing scheme

proficiency testing (3.2) 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) (3.17) 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 attribute 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 description 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 [Formula \(1\)](#).

$$x_i = \mu + \varepsilon_i \quad (1)$$

where

- x_i is the proficiency test result from participant, i ;
- μ is the true value for the measurand;
- ε_i is the 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 standard deviation of the proficiency testing results 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 within the proficiency testing scheme are considered, however, it can be 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 [Clauses 5 to 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 [Clause 11](#).

5.2 Basis of a statistical design

5.2.1 According to [ISO/IEC 17043:2010](#), 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 must include 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 must 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 must 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 must 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 must 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 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.

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

5.3.3 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. Data analysis methods 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 can be appropriate as the first step in analysis.

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

NOTE 1 Transformation of data can affect the treatment and interpretation of uncertainties associated with participant results and the assigned value.

NOTE 2 [Annex C](#) gives additional information on treatment of asymmetric outlier distributions and Example E.6 provides an example of an estimation method that is resistant to asymmetry."

5.3.4 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.

5.3.5 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 [D.1](#).

5.5 Guidelines for choosing the reporting format

5.5.1 General requirements for reporting format

5.5.1.1 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.

5.5.1.2 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 proficiency testing scheme.

NOTE 1 For some proficiency testing schemes, an objective is to evaluate a participant's ability to follow a standardized method (for example officially required or internationally standardized), 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 Reporting of replicate measurements

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 Reporting of 'less than' or 'greater than' a limit (censored data)

5.5.3.1 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.

5.5.3.2 The proficiency testing provider should either adopt validated data treatment and scoring procedures that accommodate censored data (see [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 and/or to report on whether the participant's reported upper (or lower) limit is consistent with the assigned value.

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 can result in scores that do not reflect the participant's routine service to customers.

5.5.3.3 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).

5.5.3.4 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 [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 Number of significant digits

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

5.5.4.2 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.

5.5.4.3 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.

5.5.4.4 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:

- a) The process for producing batches of the proficiency test item(s) does not change in any way that may impact homogeneity;
- b) The materials used in production of the proficiency test item(s) do not change in any way that may impact homogeneity;
- c) There is not a "failure" in homogeneity identified via either homogeneity testing or participant responses; and,
- d) 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 the standard deviation of proficiency test results 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. The procedures 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 of the proficiency testing scheme, the assessment described in [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 can then be considered sufficient to ensure adequate stability of toxic metals.

NOTE An example of homogeneity and stability checks is provided in [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 [6.4](#) to [6.6](#).

6.4 Visual review of data

6.4.1 As a first step in any data analysis the proficiency testing 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 of participant values is a useful and widely available review procedure, to look for a distribution that is unimodal and symmetric, and to identify unusual outliers (see [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 (see [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 [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 data set; 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 (MAD_e), 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 [D.2](#)).

NOTE The median, inter-quartile 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. 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 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 is best applied when the performance of participants is consistent with expectations from previous rounds of the proficiency testing scheme and there are a small number of outliers (one or two outliers on each side of the mean). Conventional tables for Grubbs' test assume a single application for a possible outlier (or 2) in a defined location, not unlimited sequential application. If Grubbs' tests are applied sequentially, the Type I error probabilities for the tests do 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 [7.3](#) to [7.7](#). The choice between these methods is the responsibility of the proficiency testing provider.

NOTE [7.3](#) to [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 [7.8](#).

7.1.3 Approaches for determining qualitative assigned values are discussed in [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 proficiency testing scheme protocol available to all participants.

7.2 Determining the uncertainty of the assigned value

7.2.1 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 [Formulae \(2\)](#) 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:

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

where

- $u(x_{\text{pt}})$ 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 (“inhomogeneity”);
- 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, can 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 7.3 to 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} can 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.

7.3.2 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.

7.3.3 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.

7.3.4 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.

7.3.5 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 [7.5](#)).

7.3.6 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 ($u(x_{\text{pt}})$) is then calculated according to [Formula \(3\)](#).

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 [Formula \(3\)](#), 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 value 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 certified property 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} sample;

$\bar{d}$ is the average of the differences d_i .

then,

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

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 [Formula \(5\)](#).

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

The example in [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 of the proficiency testing scheme.

NOTE With an assumption of a normal distribution, a difference of greater than two times the standard uncertainty can be expected to occur by chance on approximately one occasion in twenty.

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 is not sufficient

results or elsewhere suggests that the reported uncertainties are not sufficiently reliable, the consensus value should normally be obtained by the methods of [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 [6.2](#) to [6.6](#) should be used to confirm that sufficient agreement exists, before combining results.

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

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

7.7.4 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.

7.7.5 The limitations of this approach are that:

- a) there may be insufficient agreement among the participants;

- b) the consensus value may include unknown bias due to the general use of faulty methodology and this bias will not be reflected in the standard uncertainty of the assigned value;
- c) the consensus value could be biased due to the effect of bias in methods that are used to determine the assigned value.
- d) It may be difficult to determine the metrological traceability of the consensus value. While the result is always traceable to the results of the individual laboratories, a clear statement of traceability beyond that can only be made when the proficiency testing provider has complete information about the calibration standards used and control of other relevant method conditions by all of the participants contributing to the consensus value.

7.7.6 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 [E.6](#).

7.7.7 When the assigned value is derived as a robust average calculated using procedures in [C.2](#), [C.3](#), the standard uncertainty of the assigned value x_{pt} may be estimated as:

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

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. A smaller factor, or a different equation, can be justified depending on experience and the robust procedure used.

NOTE 3 An example of using an assigned value from participant results is provided in [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 u_{diff} of the difference is estimated as:

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

where

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

NOTE An example of a comparison of a reference value with a consensus value is included in [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 [7.5](#) or [7.6](#);
- 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 of the proficiency testing scheme. 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 [9.4](#) to [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 proficiency testing schemes are:

- a) performance scores have a consistent interpretation in terms of fitness for purpose from one round to the next of a proficiency testing scheme;
- 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 a proficiency testing scheme 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 of the proficiency testing scheme 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 (see 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 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 should 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 can be used to derive the following expression for the reproducibility standard deviation:

$$\sigma_R = \begin{cases} 0,22c & \text{when } c < 1,2 \times 10^{-7} \\ 0,02c^{0,849\ 5} & \text{when } 1,2 \times 10^{-7} \leq c \leq 0,138 \\ 0,01c^{0,5} & \text{when } c > 0,138 \end{cases} \quad (8)$$

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

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 [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:

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

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

NOTE This formula 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 shall have been conducted according to the requirements of [ISO 5725-2](#) or an equivalent procedure.

NOTE An example is presented in [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 proficiency testing schemes that use δ_E and D or $D\%$, P_A (Percentage of Allowed Deviation) can be used as a standardized score, for comparison across measurands and rounds (see [9.3.6](#)).