INTERNATIONAL STANDARD

ISO 13528

Third edition
2022-08

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)

Contents

Page

Foreword

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

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

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

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

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

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

This third edition of ISO 13528 cancels and replaces the second edition (ISO 13528:2015), of which it constitutes an 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.

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).

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

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 $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ 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 $\sigma_{\mathrm{pt}} = \delta_{\mathrm{E}}/3$; 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.