DRAFT INTERNATIONAL STANDARD ISO/DIS 13528

ISO/TC **60**/SC **6** Secretariat: **JISC**

Voting begins on **2013-01-22** Voting terminates on **2013-04-22**

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION • МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ • ORGANISATION INTERNATIONALE DE NORMALISATION

Statistical methods for use in proficiency testing by interlaboratory comparisons

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

[Revision of first edition (ISO 13528:2005)]

ICS 03.120.30

To expedite distribution, this document is circulated as received from the committee secretariat. ISO Central Secretariat work of editing and text composition will be undertaken at publication stage.

Pour accélérer la distribution, le présent document est distribué tel qu'il est parvenu du secrétariat du comité. Le travail de rédaction et de composition de texte sera effectué au Secrétariat central de l'ISO au stade de publication.

Contents

Page

Foreword

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

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

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

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

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

0 Introduction

0.1 The purposes for proficiency testing

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

The statistical design and analytical techniques applied must be appropriate for the stated purpose(s). This Standard presents statistical techniques for the described purposes that are consistent with other International Standards, particularly those of TC69 SC6, notably the ISO 5725 series of standards on *Accuracy: trueness and precision*. The techniques are also intended to reflect techniques from other international standards, where appropriate, and are intended to be consistent with ISO Guide 98 and JCGM 200 (VIM).

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

ISO 13528 also applies guidance from a harmonized protocol for the proficiency testing of chemical analytical laboratories [20], but is intended for use with all measurement methods. This revision of ISO 13528:2005 contains most of the information from the first edition, extended as necessary by the previously referenced documents and the extended scope of ISO/IEC 17043, which includes proficiency testing for inspection, and Annex B, which includes considerations for qualitative and ordinal results. Some procedures from the first edition have not been carried forward to this edition, due to experience which indicates those techniques are no longer considered to be appropriate.

0.2 ISO/IEC 17043

This Standard will provide support to the implementation of ISO/IEC 17043 which describes different types of proficiency testing schemes and gives guidance on the organization and design of proficiency testing schemes. Annex C of ISO/IEC 17043 gives guidance on the selection and use of proficiency testing schemes by laboratories, inspection bodies, accreditation bodies, and other interested parties. ISO/IEC 17043 should be consulted for detailed information in those areas, because the information is not duplicated here.

Annex B of ISO/IEC 17043 briefly describes the general statistical methods that are used in proficiency testing schemes. This International Standard is intended to be complementary to Annex B, providing detailed guidance that is lacking in that document on particular statistical methods for proficiency testing.

0.3 Statistical expertise

ISO/IEC 17043:2010 requires that in order to be competent, a proficiency testing provider shall have access to statistical expertise (clause 4.4.1.4) and shall authorize specific personnel to conduct statistical analysis (clause 4.2.4). Neither ISO/IEC 17043 nor this document can specify further what that necessary expertise is. For some applications an advanced degree in statistics is useful, but usually the needs for expertise can be met by individuals with technical expertise in other areas, if they are familiar with basic statistical concepts and common techniques. If an individual is charged with statistical design and/or analysis, it is very important that this person has experience with interlaboratory comparisons, even if they have 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. The guidance in

this International Standard cannot provide all the necessary expertise to consider all applications, and cannot replace the experience gained by working with interlaboratory comparisons.

0.4 Computer software

Computer software that is needed for statistical analysis can vary greatly, ranging from no software at all (for example, for proficiency testing schemes for individual participants, using known reference values and fitness for purpose evaluation criteria), to sophisticated statistical packages used for proficiency testing schemes with large numbers of participants and measurands, or using complicated robust methods. Most of the techniques in this International Standard can be accomplished by conventional spread sheet applications, perhaps with customised analysis for a particular scheme or analysis; some techniques will require computer applications that are freely available (at the time of publication). In all cases, the user must verify the accuracy of their calculations, especially when special routines have been entered by the user. However, even when the techniques in this International Standard are appropriate and correctly implemented by adequate computer applications, they cannot be applied without experienced oversight by an individual with technical and statistical expertise that is sufficient to identify and investigate anomalies that can occur in any round of proficiency testing.

Statistical methods for use in proficiency testing

1 Scope

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

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

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

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

2 Normative references

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

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

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

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

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

JCGM 200:2012, International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 3rd edition

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

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-2, ISO 5725-1, ISO/IEC 17043, JCGM 200, ISO Guide 34, and the following apply.

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

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

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

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

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

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

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

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

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

g) sampling — where samples are taken for subsequent analysis; and

h) data transformation and 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, based on the available information

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

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

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

NOTE 4 The notes for the definition in this document are not the same as the definition in ISO/IEC 17043.

3.5
***z* score**
Standardized measure of performance, calculated using the participant result, assigned value and the standard deviation for proficiency assessment

3.6
outlier
Observation in a set of data that appears to be inconsistent with the remainder of that set

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

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

3.7
participant
Laboratory, organization, or individual that receives proficiency test items and submits results for review by the proficiency testing provider.

3.8
proficiency test item
Sample, product, artefact, reference material, piece of equipment, measurement standard, data set or other information used for proficiency testing.

NOTE In most instances, proficiency test items meet the ISO Guide 34 definition of “reference material” (3.11 below)

3.9
proficiency testing provider
Organization which takes responsibility for all tasks in the development and operation of a proficiency testing scheme.

3.10
proficiency testing scheme
proficiency testing designed and operated in one or more rounds for a specified area of testing, measurement, calibration or inspection.

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

3.11
reference material (RM)
Material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process. (ISO Guide 34:2009)

NOTE 1 RM is a generic term

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

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

3.12
certified reference material (CRM)
Reference material characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability. (ISO Guide 34:2009)

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

4 Guidelines for the statistical design of proficiency testing schemes

4.1 Introduction

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

The following sections give guidance on the design of proficiency testing schemes and on the statistical treatment of results, including the calculation and interpretation of various performance scores.

4.2 Basis of a statistical design

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

NOTE For example, schemes with the following main approaches could have different statistical designs:

— a scheme to compare participant results with combined results from a group in the same round, and limits determined by the variability of participant results;

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

— a scheme to compare participants' results with the participant's own measurement uncertainty.

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

4.2.2 There are various types of data used in proficiency testing, including quantitative, nominal (categorical), and ordinal. Among the continuous variables, some results might be on a differential or rational scale (a scale with an arbitrary 0); or a relative, or ratio scale (on a scale with a true 0). For some measurements on a continuous scale, only a discrete and discontinuous set of values can be realized; however, in many cases these results can be treated as other continuous variables.

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

4.3 Considerations for the statistical distribution of results

4.3.1 Statistical assumption for distribution

4.3.1.1 The design and data analysis techniques shall be consistent with the statistical assumptions for the data (see ISO/IEC 17043 clause 4.4.4.2). Most common analysis techniques for proficiency testing assume that a set of results from competent participants will be unimodal, reasonably symmetric, and approximately normally distributed. A common additional assumption is that the distribution of results from all participants is mixed (or 'contaminated') with results from a population of erroneous values and outliers. Usually, the scoring interpretation relies on the assumption of normality, but only for the underlying assumed distribution for competent participants.

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

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

— transformation to provide approximate symmetry;

— methods of estimation that are resistant to asymmetry;

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

NOTE For example, dilution based results, such as for quantitative microbiological counts or for immunoassay techniques, are often distributed according to the logarithmic normal distribution, and so a logarithmic transformation is appropriate as the first step in analysis. Another example is counts of small numbers of particles, which may be distributed according to a Poisson distribution, and therefore the evaluation interval would be determined using a table of Poisson probabilities, based on the average count for the group of participants.

4.3.1.4 Many calibration measurands follow defined statistical distributions that are described in the measurement procedure; these distributions should be considered in any evaluation protocol.

4.3.2 Reasonableness of assumptions

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

NOTE 1 The demonstration of the reasonableness of a distribution assumption is less rigorous than the demonstration of the validity of that assumption.

NOTE 2 In general, it is not appropriate to determine evaluations using an objective to classify a certain proportion of results as generating an 'action signal', based on an assumed distribution (e.g., 'action signal' = outside 95% confidence limits for the normal distribution).

4.4 Considerations for small numbers of participants

4.4.1 Appropriate methods

Statistical methods that are appropriate for large numbers of participants may not be appropriate with limited numbers of participants. The appropriate number of participants cannot be specified exactly for all cases since it will vary for different summary statistics, the comparability of results from different methods, the experience of the participants, and the experience of the provider. For example, reliable means and medians can usually be determined with fewer results than can standard deviations, and the necessary numbers of results are usually smaller when participants have experience with proficiency testing, than when a scheme is new or there is a large number of inexperienced participants.

4.4.2 Alternatives for small numbers

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

a) Visually review results to assure sufficient agreement among the limited group;

b) Review data from similar items and measurands used in previous rounds or other schemes, and use evaluation criteria from those rounds or schemes. Calculate the expanded uncertainty of the assigned value using the appropriate expansion factor from the t distribution, and assess whether the expanded uncertainty U_X is < $0.6\,\hat{\sigma}$ (similar to u_X < $0.3\,\hat{\sigma}$).

NOTE When there are few participants in a scheme, the recommendations of the IUPAC/CITAC Technical Report: *Selection and use of proficiency testing schemes for a limited number of participants* should be considered, if feasible. 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 measurement 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.

4.5 Considerations for qualitative data (including nominal and ordinal properties)

4.5.1 Types of qualitative data

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

— Nominal properties, where the property value has no magnitude (such as a type of substance or organism).