# GUIDE 43-1

# Proficiency testing by interlaboratory comparisons —

# Part 1: Development and operation of proficiency testing schemes



Second edition 1997

# Contents







International Organization for Standardization
Case postale 56 • CH-1211 Genève 20 • Switzerland
Internet central@iso.ch
X.400 c=ch; a=400net; p=iso; o=isocs; s=central

Printed in Switzerland

## Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

ISO/IEC Guide 43-1 was prepared by ISO/CASCO Ad Hoc Group for Revision of ISO/IEC Guide 43. A draft was circulated to CASCO members and IEC National Committees for comments. A final draft has subsequently been approved by ISO/CASCO and by IEC Council for publication as an ISO/IEC Guide.

Parts 1 and 2 of ISO/IEC Guide 43 cancel and replace the first edition (ISO/IEC Guide 43:1984).

ISO/IEC Guide 43:1984 covered guidance on development and operation of laboratory proficiency testing with limited emphasis on the use of the outcomes of proficiency testing by accreditation bodies. It is now intended to provide guidance in three areas, namely:

a) to distinguish between use of interlaboratory comparisons for proficiency testing and for other purposes;

b) the development and operation of interlaboratory comparisons for use in proficiency testing schemes; and

c) the selection and use of proficiency testing schemes by laboratory accreditation bodies.

ISO/IEC Guide 43 consists of the following parts, under the general title *Proficiency testing by interlaboratory comparisons:*

— *Part 1: Development and operation of proficiency testing schemes*

— *Part 2: Selection and use of proficiency testing schemes by laboratory accreditation bodies*

Annexes to this part of ISO/IEC Guide 43 provide statistical guidance on treatment of data from proficiency testing schemes and guidelines on documentation (Quality Manual) for the operation of proficiency testing schemes.

# Introduction

Interlaboratory comparisons are conducted for a number of purposes and may be used by participating laboratories and other parties.

Interlaboratory comparisons may be used, for example, to:

a) determine the performance of individual laboratories for specific tests or measurements and to monitor laboratories' continuing performance;

b) identify problems in laboratories and initiate remedial actions which may be related to, for example, individual staff performance or calibration of instrumentation;

c) establish the effectiveness and comparability of new test or measurement methods and similarly to monitor established methods;

d) provide additional confidence to laboratory clients;

e) identify interlaboratory differences;

f) determine the performance characteristics of a method — often known as collaborative trials;

g) assign values to reference materials (RMs) and assess their suitability for use in specific test or measurement procedures.

*Proficiency testing* is the use of interlaboratory comparisons for purpose a); i.e. the determination of laboratory testing or measurement performance. However, the operation of proficiency testing schemes may often also provide information for the other purposes listed above.

Participation in proficiency testing schemes provides laboratories with an objective means of assessing and demonstrating the reliability of the data they are producing. Although there are several types of proficiency testing schemes (see clause 4), most share the common feature of the comparison of test and measurement results obtained by two or more laboratories.

One of the main uses of proficiency testing schemes is to assess laboratories' ability to perform tests competently. This may include assessment by laboratories themselves, by their clients, or by other parties such as accreditation or regulatory bodies. It thus supplements laboratories' own internal quality control procedures by providing an additional external measure of their testing capability. These activities also complement the technique of on-site laboratory assessment by technical specialists (usually used by laboratory accreditation bodies). Confidence that a testing or calibration laboratory consistently obtains reliable results is of major importance to users of laboratory services. Users seeking such an assurance may undertake their own evaluation of results or may use the evaluation of other bodies.

While the emphasis of this part of ISO/IEC Guide 43 is on operation of interlaboratory comparisons for proficiency testing, most of the principles and guidance given are applicable to operation of interlaboratory comparisons for other purposes.

While many laboratory accreditation bodies operate their own proficiency testing schemes, a significant number also use proficiency testing schemes or other forms of interlaboratory comparisons operated by other bodies. The purpose of part 2 of ISO/IEC Guide 43 is to provide harmonized principles for the selection of suitable interlaboratory comparisons for use as proficiency testing schemes by laboratory accreditation bodies.

Most bodies assessing the technical competence of laboratories require or expect satisfactory performance in proficiency testing schemes as significant evidence of a laboratory's ability to produce reliable results (except where proficiency testing is inappropriate).

However, it is emphasized that a major distinction exists between:

a) the evaluation of the competence of a laboratory by the assessment of its total operation against predetermined requirements; and

b) the examination of the results of a laboratory's participation in proficiency testing which may only be considered as giving information about the technical competence of the testing laboratory at a single point of time under the specific conditions of the test (or tests) involved in a particular proficiency testing scheme.

In preparing this Guide, reference was made to a number of guidance documents relevant to proficiency testing produced by ILAC; ISO (TC 69); ISO/REMCO; IUPAC; AOAC; ASTM; and WECC and WELAC (now combined as EAL).

# Proficiency testing by interlaboratory comparisons —
# Part 1:
# Development and operation of proficiency testing schemes



## 1 Scope

While there are a number of uses of interlaboratory comparisons, and variations in their design and implementation, it is still possible to specify the essential principles that need to be considered when organizing such comparisons. This part of ISO/IEC Guide 43 defines those principles and describes the factors which should be taken into account in the organization and conduct of proficiency testing schemes.

Part 2 of ISO/IEC Guide 43 describes how laboratory accreditation bodies, who assess technical competence of laboratories, should select and use proficiency testing schemes.

This part of ISO/IEC Guide 43 is intended for use by proficiency testing operators and users such as participant laboratories, accreditation bodies, regulatory authorities and clients of laboratory services who have a need to assess the technical competence of laboratories. It is particularly useful for laboratories in self-evaluation, but recognizes that proficiency testing is only one mechanism which can contribute to the establishment of mutual confidence between users of different testing laboratories.

It is currently a condition of some accreditation bodies that laboratories participate regularly in proficiency testing schemes that they have accepted as fit for purpose. Therefore, it is essential that operators of such schemes comply with principles for conduct of professionally managed proficiency schemes, both in terms of technical requirements, statistical procedures (see examples in annex A), and in quality management (see guidance in annex B).

The methods of operation within different proficiency testing organizations are not expected to be identical and this Guide does not give specific operational details for interlaboratory comparisons. The contents of this Guide are intended only as a framework to be modified appropriately for particular situations, including schemes with either small or large numbers of participants.

This Guide is not intended to cover a technique often used by organizations to evaluate a single laboratory's performance through submissions of certified reference materials or other well-characterized test items.

A bibliography is given in annex C.

## 2 References

ISO 3534-1:1993, *Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms.*

ISO 5725-1:1994, *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.*

ISO 5725-2:1994, *Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.*

ISO 5725-4:1994, *Accuracy (trueness and precision) of measurement methods and results — Part 4: Basic methods for the determination of the trueness of a standard measurement method.*

*ISO 9000 Quality Management Compendium*, 1994.

ISO/IEC Guide 2:1996, *General terms and their definitions concerning standardization and related activities.*

ISO/IEC Guide 25:1990, *General requirements for the competence of calibration and testing laboratories.*

ISO/IEC Guide 43-2:1997, *Proficiency testing by interlaboratory comparisons — Part 2: Selection and use of*

*proficiency testing schemes by laboratory accreditation bodies.*

*Guide to the Expression of Uncertainty in Measurement: 1993*, issued by BIPM; IEC; IFCC; ISO; IUPAC; IUPAP; OIML.

*International vocabulary of basic and general terms in metrology*: 1993, issued by BIPM; IEC; IFCC; ISO; IUPAC; IUPAP; OIML. [VIM:1993]

The International Harmonized Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories. *Journal of AOAC International*, **76**, No. 4, 1993, pp. 926-940.

*Evaluation of Matrix Effects: Proposed Guideline*, NCCLS Document EP-14P. National Committee for Clinical Laboratory Standards, Villanova, PA, 1994.

## 3 Definitions

For the purposes of this Guide, the following definitions apply. Some definitions are taken from other ISO Guides and International Standards, as shown.

**3.1**
**test**
technical operation that consists of the determination of one or more characteristics of a given product, process or service according to a specified procedure

[ISO/IEC Guide 2]

**3.2**
**testing laboratory**
laboratory that performs tests

NOTE — The term "testing laboratory" can be used in the sense of a legal entity, a technical entity or both.

[ISO/IEC Guide 2]

**3.3**
**test item**
material or artefact presented to the participating laboratory for the purpose of proficiency testing

**3.4**
**test method**
specified technical procedure for performing a test

[ISO/IEC Guide 2]

**3.5**
**test result**
the value of a characteristic obtained by completely carrying out a specified measurement method

[ISO 5725-1]

**3.6**
**(laboratory) proficiency testing**
determination of laboratory testing performance by means of interlaboratory comparisons

[ISO/IEC Guide 2]

NOTE — For the purposes of this Guide, the term laboratory proficiency testing is taken in its widest sense and includes, for example:

a) Qualitative schemes — for example where laboratories are required to identify a component of a test item.

b) Data transformation exercises — for example where laboratories are furnished with sets of data and are required to manipulate the data to provide further information.

c) Single item testing — where one item is sent to a number of laboratories sequentially and returned to the organizer at intervals.

d) One-off exercises — where laboratories are provided with a test item on a single occasion.

e) Continuous schemes — where laboratories are provided with test items at regular intervals on a continuing basis.

f) Sampling — for example where individuals or organizations are required to take samples for subsequent analysis.

**3.7**
**interlaboratory comparisons**
organization, performance and evaluation of tests on the same or similar test items by two or more laboratories in accordance with predetermined conditions

NOTE — In some circumstances, one of the laboratories involved in the intercomparison may be the laboratory which provided the assigned value for the test item.

**3.8**
**reference material (RM)**
material or substance one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to the materials

[VIM:1993, 6.13]

**3.9**
**certified reference material (CRM)**
reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes its traceability to an accurate realization of the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence

[VIM:1993, 6.14]

**3.10**
**reference laboratory**
laboratory that provides reference values on a test item

NOTE — For example, a National Calibration Laboratory.

**3.11**
**assigned value**
value attributed to a particular quantity and accepted, sometimes by convention, as having an uncertainty appropriate for a given purpose

[see VIM:1993, 1.20 and notes 1 and 2]

**3.12**
**traceability**
property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties

[VIM:1993, 6.10]

**3.13**
**coordinator**
organization (or person) with responsibility for co-ordinating all of the activities involved in the operation of a proficiency testing scheme

**3.14**
**trueness**
closeness of agreement between the average value obtained from a large series of test results and an accepted reference value

[ISO 3534-1]

**3.15**
**precision**
closeness of agreement between independent test results obtained under prescribed (stipulated) conditions

[ISO 3534-1]

**3.16**
**outlier**
member of a set of values which is inconsistent with the other members of that set

[ISO 5725-1]

**3.17**
**extreme results**
outliers and other values which are grossly inconsistent with other members of the data set

NOTE — These results can have a profound infuence on summary statistics such as the mean and standard deviation.

**3.18**
**robust statistical techniques**
techniques to minimize the influence that extreme results can have on estimates of the mean and standard deviation

NOTE — These techniques assign less weight to extreme results, rather than eliminate them from a data set.

**3.19**
**uncertainty of measurement**
parameter associated with the results of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand

[VIM:1993, 3.9]



## 4 Types of proficiency testing

### 4.1 General

Proficiency testing techniques vary depending on the nature of the test item, the method in use and the number of laboratories participating. Most possess the common feature of comparison of results obtained by one laboratory with those obtained by one or more other laboratories. In some schemes, one of the participating laboratories may have a controlling, co-ordinating, or reference function.

The following are common types of proficiency testing schemes.

### 4.2 Measurement comparison schemes

Measurement comparison schemes involve the test item to be measured or calibrated being circulated successively from one participating laboratory to the

next. Features of such schemes usually are as follows.

a) Assigned values for the test item are provided by a Reference Laboratory, which might be a country's highest authority for the measurement concerned. It may be necessary for the test item to be checked at specific stages during the conduct of the proficiency test. This is to ensure that there are no significant changes in the assigned value throughout the course of the proficiency test.

b) Schemes involving sequential participation take time (in some cases years) to complete. This causes a number of difficulties such as: ensuring the stability of the item; the strict monitoring of its circulation and the time allowed for measurement by individual participants; and the need to supply feedback on individual performance to laboratories during the scheme's implementation, rather than waiting until it finishes. In addition, it may be difficult to compare results on a group basis as there may be relatively few laboratories whose measurement capabilities closely match each other.

c) The individual measurement results are compared with the reference values established by the Reference Laboratory. The coordinator should take into account the claimed measurement uncertainty of each participating laboratory.

d) Examples of items (measurement artefacts) used in this type of proficiency testing include reference standards (e.g. resistors, gauges and instruments).

## 4.3 Interlaboratory testing schemes

Interlaboratory testing schemes involve randomly selected sub-samples from a source of material being distributed simultaneously to participating testing laboratories for concurrent testing. Occasionally, this technique is also used for interlaboratory measurement schemes. After completion of the testing, the results are returned to the coordinating body, and compared with the assigned value(s) to give an indication of the performance of the individual laboratories and the group as a whole.

Examples of test items used in this type of proficiency testing include food, body fluids, water, soils and other environmental material. In some cases, separate portions of previously established (certified) reference materials are circulated.

It is essential that the batch of test items provided to participants in each round be sufficiently homogeneous so that any results later identified as extreme are not attributed to any significant test item variability. (See 5.6.2 and A.4 in annex A.)

Interlaboratory testing type schemes are commonly used by accreditation bodies, regulatory bodies and other organizations when they utilize schemes in the testing field.

One common interlaboratory test scheme is the "split-level" design where similar (but not identical) levels of measurand are included in two separate test items. This design is used to estimate laboratory precision at a specific level of a measurand. It avoids problems associated with replicate measurements on the same test item, or with inclusion of two identical test items in the same proficiency test round.

## 4.4 Split-sample testing schemes

One special form of proficiency testing which is often used by clients of laboratories, including some regulatory bodies, is the technique of split-sample testing. (This technique should not be confused with split-level schemes which are discussed in 4.3.)

Typically, split-sample testing involves comparisons of the data produced by small groups of laboratories (often only two laboratories) which are being evaluated as potential, or continuing, suppliers of testing services.

Similar intercomparisons are regularly conducted in commercial transactions when samples representing a traded commodity are split between a laboratory representing the supplier and another laboratory representing the purchaser. An additional sample is normally retained for a third-party laboratory to test if arbitration is needed on any significant differences between the results produced by the supplier's and the purchaser's laboratories.

Split-sample testing schemes involve samples of a product or a material being divided into two or more parts with each participating laboratory testing one part of each sample. They differ from the type of proficiency testing described in 4.3, as there is usually a very limited number of participating laboratories (often two). Uses for this type of scheme include identifying poor precision, describing consistent bias and verifying the effectiveness of corrective actions.

Such schemes often need retention of sufficient material to resolve any perceived differences between the limited number of laboratories involved by further analysis by additional laboratories.

A similar technique of split-sample testing is also used in the monitoring of clinical and environmental laboratories. Typically, these schemes involve the results from several split samples over a wide concentration interval being compared between an individual laboratory and one or more other laboratories. Under such schemes, one of the laboratories may be considered to operate at a higher metrological level (i.e. lower level of uncertainty) due to the use of reference methodology and more advanced equipment, etc. Its results are considered to be the reference values in such intercomparisons and it may act as an advisory or mentor laboratory to the other laboratories comparing split-sample data with it.

### 4.5 Qualitative schemes

Evaluation of laboratory testing performance will not always involve interlaboratory comparisons. [See a) in note to 3.6.] For example, some schemes are designed to evaluate the capabilities of laboratories to characterize specific entities (e.g. type of asbestos, identity of a specific pathogenic organism, etc.).

Such schemes may involve the special preparation of test items with addition of the subject component by the scheme coordinator. As such, the schemes are "qualitative" in nature, and do not need the involvement of multiple laboratories or interlaboratory comparisons to evaluate a laboratory's testing performance.

### 4.6 Known-value schemes

Other special types of proficiency testing schemes may involve the preparation of test items with known amounts of the measurand under test. It is then possible to evaluate the capability of an individual laboratory to test the item and provide numerical results for comparison with the assigned value. Once again, such proficiency schemes do not need the involvement of multiple laboratories.

### 4.7 Partial-process schemes

Special types of proficiency testing involve the evaluation of laboratories' abilities to perform parts of the overall testing or measurement process. For example, some existing proficiency schemes evaluate laboratories' abilities to transform and report a given set of data (rather than conduct the actual test or measurement) or to take and prepare samples or specimens in accordance with a specification.

## 5 Organization and design

### 5.1 Framework

**5.1.1** The design stage of any proficiency testing scheme requires the input of technical experts, statisticians and a scheme coordinator to ensure its success and smooth operation.

**5.1.2** The coordinator, in consultation with these other personnel, should develop a scheme appropriate to the particular proficiency test. A proficiency test scheme should be designed to avoid any confusion about its objectives. A plan should be agreed upon and documented (see annex B) before the start of the scheme and typically would include the following information:

a) the name and the address of the organization conducting the proficiency scheme;

b) the name and address of the coordinator and other personnel involved in the design and operation of the proficiency scheme;

c) the nature and the purpose of the proficiency scheme;

d) a procedure for the manner in which the participants are selected, where appropriate, or criteria which need to be met before participation is allowed;

e) the name and address of the laboratory or laboratories performing (parts of) the scheme (e.g. sampling, sample processing, homogeneity testing and assigning values), and the number of expected participants;

f) the nature of the test item(s) and test(s) selected, as well as a short description of the considerations underlying these choices;

g) a description of the manner in which the test items are obtained, processed, checked and transported;

h) a description of the information that is supplied to participants in this notification phase and of the time schedule for the various phases of the proficiency testing;

i) the expected initial and target dates or deadlines of the proficiency scheme including the date(s) for the testing to be carried out by the participants;

j) for on-going schemes, the frequency at which test items are distributed;

k) information on methods or procedures which participants may need to use to perform the tests or measurements (commonly their routine procedures);

l) an outline of the statistical analysis to be used, including the determination of assigned value(s) and any outlier detection techniques;

m) a description of the data or information to be returned to participants;

n) the basis for performance evaluation techniques;

o) a description of the extent to which the test results, and the conclusions that will be based on the outcome of the proficiency tests, are to be made public.

## 5.2 Staff

**5.2.1** The staff involved in providing the scheme should have, or collaborate closely with those holding adequate qualifications and experience in the design, implementation and reporting of interlaboratory comparisons. They should include appropriate technical, statistical and administrative skills.

**5.2.2** As mentioned in 5.1.1, the operation of particular interlaboratory comparisons will also require the guidance of persons with detailed technical knowledge and experience of the test methods and procedures involved. To this end the coordinator may need to enlist one or more appropriate persons drawn from, for example, professional bodies, a contract laboratory (if any), scheme participants or end users of the data, to act as an advisory group.

**5.2.3** The functions of this advisory group may include:

a) the development and review of procedures for the planning, execution, analysis, reporting and effectiveness of the proficiency testing scheme;

b) the identification and evaluation of interlaboratory comparisons organized by other bodies;

c) the evaluation of proficiency test results regarding the performance of participating laboratories;

d) providing advice to any body assessing the technical competence of participating laboratories, both on the results obtained during a proficiency test scheme, and how those results should be used with other aspects of laboratory evaluations;

e) providing advice to participants apparently experiencing problems; and

f) resolving any disputes between the coordinator and participants.

## 5.3 Data-processing equipment

Whatever equipment is used, it should be adequate to conduct all necessary data entry and statistical analysis and provide timely and valid results. Procedures for checking data entry should be implemented and all software should be verified, supported and backed up. The storage and security of data files should be controlled.

## 5.4 Statistical design

**5.4.1** The statistical model and data analysis techniques to be used should be documented, together with a short description of the background to their selection. Further details of common statistical procedures and treatment of proficiency testing data are discussed in annex A.

**5.4.2** Appropriate statistical design of a proficiency testing scheme is essential. Careful consideration should be given to the following matters and their interaction:

a) the precision and trueness of the test(s) involved;

b) the smallest differences to be detected between participating laboratories at a desired confidence level;

c) the number of participating laboratories;

d) the number of samples to be tested and the number of repeat tests or measurements to be carried out on each sample;

e) the procedures to be used to estimate the assigned value;

f) procedures to be used to identify outliers.

**5.4.3** In the absence of reliable information concerning a), it may be necessary in some cases to organize a pilot interlaboratory comparison (collaborative trial) to obtain it.

## 5.5 Test item preparation

**5.5.1** Preparation of test items may either be contracted out or undertaken by the coordinator. The

organization preparing the test item should have demonstrable competence to do so.

**5.5.2** Any conditions relating to the test items which may affect the integrity of the interlaboratory comparison, such as homogeneity, sampling, stability, possible damage in transit and effects of ambient conditions, should be considered (see 5.6).

**5.5.3** The test items or materials to be distributed in the scheme should generally be similar in nature to those routinely tested by participating laboratories.

NOTE — An example of a protocol for establishing such similarity is given in document NCCLS EP-14P, published by the National Committee for Clinical Laboratory Standards, Villanova, PA, 1994.

**5.5.4** The number of test items to be distributed may depend on whether there is a requirement to cover a range of compositions.

**5.5.5** The assigned value(s) should not be disclosed to the participants until after the results have been collated. However, in some cases it may be appropriate to advise target ranges prior to testing.

**5.5.6** Consideration could be given to preparation of additional test items other than those needed for the proficiency test scheme. Surplus test items may potentially be useful as a reference material, quality control material or training aid for laboratories after results from participants have been evaluated.

## 5.6 Test item management

**5.6.1** Procedures for sampling, randomizing, transporting, receiving, identifying, labelling, storing and handling of test items should be documented.

**5.6.2** Where bulk material is prepared for a proficiency test, it should be sufficiently homogeneous for each test parameter so that all laboratories will receive test items that do not differ significantly in the parameters to be measured. The coordinator should document the procedure used to establish the homogeneity of the test item (see A.4 in annex A). When possible, homogeneity testing should be carried out prior to the despatch of the test items to the participating laboratories. The degree of homogeneity should be such that differences between test items will not significantly affect the evaluation of a participant's result.

**5.6.3** Where possible, the coordinator should also provide evidence that the test items are sufficiently stable to ensure that they will not undergo any significant change throughout the conduct of the proficiency test. When unstable measurands need to be assessed, it may be necessary for the coordinating organization to specify a date by which the testing should be completed, and any required special pretesting procedures.

**5.6.4** Coordinators should consider any hazards that the test items might pose and take appropriate action to advise any party that might be at risk (e.g. test material distributors, testing laboratories, etc.) of the potential hazard involved.

## 5.7 Choice of method/procedure

**5.7.1** Participants will normally be able to use the method of their choice, which is consistent with routine procedures used in their laboratories. However, in certain circumstances, the coordinator may instruct participants to use a specified method. Such methods are usually nationally or internationally accepted standard methods, and will have been validated by an appropriate procedure (e.g. collaborative trial).

**5.7.2** Where a calibration procedure is used, the assigned value will often be a reference value obtained from measurements obtained by a high-echelon calibration laboratory (often a National Standards Laboratory) which should use a well-defined and accepted procedure. It is desirable that participating laboratories use the same or similar procedure, but this will not always be practicable for calibration laboratories.

**5.7.3** Where participants are free to use a method of their own choice, coordinators should, where appropriate, request details of the methods used to allow the use of participants' results to compare and comment on the methods.

## 5.8 Evolution of proficiency testing schemes

To ensure that proficiency testing schemes are able to adapt to technical and scientific developments, they may need to include new types of samples or new methods or procedures. Early conclusions from the results of such schemes on the performance of individual laboratories should be drawn with due care. (See 6.4.3.)

# 6 Operation and reporting

## 6.1 Coordination and documentation

The day-to-day operation of a scheme should be the responsibility of a coordinator. All practices and procedures should be documented. These may be included in, or supplemented by, a quality manual. (See annex B.)

## 6.2 Instructions

**6.2.1** Detailed instructions covering all aspects of the scheme which should be adhered to by the participating laboratories should be provided. These may be provided, for example, as an integral part of a scheme protocol.

**6.2.2** Instructions may include details concerning factors which could influence the testing of the supplied test items or materials. Such factors may include operators, the nature of items or materials, equipment status, selection of test procedure and timing of testing.

**6.2.3** Specific instructions on the recording and reporting of test or calibration results may also be supplied (e.g. units, number of significant figures, reporting basis, result deadlines, etc.).

**6.2.4** Participants should be advised to treat proficiency test items as if they were performing routine tests (unless there are some special requirements in the design of the proficiency test which may require departure from this principle).

## 6.3 Packaging and transportation

The coordinator of the scheme should consider the following aspects regarding the distribution of the test or measurement item. The packaging and method of transport have to be adequate and able to protect the stability and characteristics of the test items. There may be certain restrictions on transportation such as dangerous goods regulations or customs requirements. In some cases, the laboratories themselves should also take responsibility for the transport of the items, particularly in sequential measurement comparison schemes.

All appropriate customs declaration forms should be completed by the coordinator to ensure that delays in customs clearance are minimized. The scheme will need to comply with national and international regulations applicable to test item transport.

## 6.4 Data analysis and records

**6.4.1** The results received from the participating laboratories should be entered and analysed, then reported back as soon as practicable. It is essential that procedures are put in place to check the validity of data entry and transfers and subsequent statistical analysis (see 5.3). It is recommended that data sheets, computer back-up files, printouts, graphs, etc. be retained for a specified period.

**6.4.2** Data analysis should generate summary measures and performance statistics and associated information consistent with the scheme's statistical model and the objectives of the scheme. The influence of extreme results on summary statistics should be minimized by the use of outlier detection tests to identify and then omit them or, preferably, by the use of robust statistics. Annex A contains some broad suggestions for statistical evaluations.

**6.4.3** Scheme coordinators should have documented criteria for dealing with test results that may be inappropriate for proficiency evaluations. For example, it is recommended that for measurands for which the test material has been shown not to be sufficiently homogeneous or stable for the purposes of a proficiency test, no grading or scoring should be given.

## 6.5 Scheme reports

**6.5.1** The content of scheme reports will vary depending on the purpose of a particular scheme, but should be clear and comprehensive and include data on the distribution of results from all laboratories together with an indication of individual participant's performance (see 6.6).

**6.5.2** The following information should normally be included in reports of proficiency schemes:

a) name and address of the organization conducting or coordinating the scheme;

b) names and affiliations of persons involved in the design and conduct of the scheme (see 5.2);

c) date of issue of report;

d) report number and clear identification of scheme;

e) clear description of items or materials used, including details of sample preparation and homogeneity testing;

f) laboratory participation codes and test results;

g) statistical data and summaries, including assigned values and range of acceptable results;

h) procedures used to establish any assigned value;

i) details of the traceability and uncertainty of any assigned value;

j) assigned values and summary statistics for test methods/procedures used by other participating laboratories (if different methods are used by different laboratories);

k) comments on laboratory performance by the coordinator and technical advisers (see 6.6);

l) procedures used to design and implement the scheme (which may include reference to a scheme protocol);

m) procedures used to statistically analyse the data (see annex A);

n) advice, where appropriate, on the interpretation of the statistical analysis.

**6.5.3** For schemes operated on a regular basis, it may be sufficient to have simpler reports such that many of the recommended elements in 6.5.2 could be excluded from routine reports, but included in periodic summary reports and on request from participants.

**6.5.4** Reports should be made available quickly within specified timetables. Although, ideally, all original data supplied should be reported to participants, it may not be possible to achieve this in some very extensive schemes. Participants should receive at least the results of all laboratories in summary (e.g. graphical) form. In some schemes such as long period measurement comparison schemes, interim reports should be issued to individual participants.

## 6.6 Evaluation of performance

**6.6.1** Where an evaluation of performance is needed the coordinator should be responsible for ensuring that the method of evaluation is appropriate to maintain the credibility of the scheme.

**6.6.2** The coordinator may enlist the assistance of technical advisers to provide expert commentary on performance with respect to:

a) overall performance versus prior expectations (taking uncertainties into account);

b) variation within and between laboratories (and comparisons with any previous schemes or published precision data);

c) variation between methods or procedures, if applicable;

d) possible sources of error (refer extreme results) and suggestions for improving performance;

e) any other suggestions, recommendations or general comments;

f) conclusions.

**6.6.3** It may be necessary to provide individual summary sheets for participants periodically during or after a particular scheme. These may include updated summaries of performance of individual laboratories over various rounds of an on-going scheme. Such summaries can be further analysed and trends highlighted if required.

**6.6.4** A variety of procedures exist to assess performance of participants, both for one-off schemes and also after consecutive rounds of on-going schemes. Some examples of procedures are given in annex A.

**6.6.5** Reporting of performance by ranking laboratories in a table according to their performance is not recommended in proficiency testing. Therefore, ranking should only be used with extreme caution as it can be misleading and open to misinterpretation.

## 6.7 Communication with participants

**6.7.1** Participants should be provided with a detailed set of information on joining a proficiency testing scheme, such as a formal scheme protocol. Subsequent communication with participants can be by letter, newsletter and/or reports, together with periodic open meetings. Participants should be advised immmediately of any changes in scheme design or operation.

**6.7.2** Participants should be able to refer to the coordinator if they consider that assessment of their performance in a proficiency test is in error.

**6.7.3** Feedback from laboratories should be encouraged, so that participants actively contribute to the development of a scheme.

**6.7.4** The procedures associated with the corrective action undertaken by participants (particularly in relation to feedback to accreditation bodies) is addressed in part 2 of ISO/IEC Guide 43.

# 7 Confidentiality/ethical considerations

## 7.1 Confidentiality of records

Normally, it is the policy of most schemes to maintain confidentiality of the identity of individual participants. The identity of participants should only be known to the minimum number of people involved in coordinating a programme, and this should extend to any subsequent remedial advice or action applied to a laboratory exhibiting poor performance. In some circumstances, a coordinating body may be required to report poor performance to a particular authority, but participants should be notified of this possibility when agreeing to participate in the scheme.

A group of participants may elect to waive confidentiality within the group, for the purposes of discussion and mutual assistance in improvement.

## 7.2 Collusion and falsification of results

Although proficiency testing schemes are intended primarily to help participants improve their performance, there may be a tendency among some participants to provide a falsely optimistic impression of their capabilities. For example, collusion may take place between laboratories, so that truly independent data are not submitted. Laboratories may also give a false impression of their performance if they routinely carry out single analyses, but report the mean of replicate determinations on the proficiency test items or conduct additional replicates to those specified for a particular scheme. Proficiency testing schemes should, where practicable, be designed to ensure that there is as little collusion and falsification as possible.

Although all reasonable measures should be taken by the coordinators to prevent collusion, it should be appreciated that it is the responsibility of the participating laboratories to avoid it.

# Annex A
(informative)

# Examples of statistical methods for treatment of proficiency test data

Proficiency test results can appear in many forms, spanning a wide range of data types and underlying statistical distributions. The statistical techniques used to analyse the results need to be appropriate for each situation, and so are too varied to specify.

There are, however, three steps common to all proficiency tests, when participants' results are to be evaluated:

a) determination of the assigned value;

b) calculation of performance statistics;

c) evaluation of performance;

and, in some cases,

d) preliminary determination of test item homogeneity and stability.

This annex gives general criteria for statistical techniques that can be applied as needed to guide specific applications.

With new interlaboratory comparison schemes, agreement initially is often poor due to new questions, new forms, artificial test items, poor agreement of methods, or variable laboratory procedures. Coordinators may have to use robust measures of relative performance (such as percentiles) until agreement improves. Statistical techniques may need to be refined once interlaboratory agreement has improved and proficiency testing is well established.

This annex does not consider statistical techniques for analytical studies other than for treatment of proficiency test data. Different techniques may be needed to implement the other uses of interlaboratory comparison data listed in the Introduction.

NOTE — ISO/TC 69 is currently preparing a document providing detailed information on statistical methods contained in this annex.

## A.1 Determination of the assigned value and its uncertainty

**A.1.1** There are various procedures available for the establishment of assigned values. The most common procedures are listed below in an order that, in most cases, will result in increasing uncertainty for the assigned value. These procedures involve use of:

a) **Known values** — with results determined by specific test item formulation (e.g. manufacture or dilution).

b) **Certified reference values** — as determined by definitive methods (for quantitative tests).

c) **Reference values** — as determined by analysis, measurement or comparison of the test item alongside a reference material or standard, traceable to a national or international standard.

d) **Consensus values from expert laboratories** — expert laboratories should have demonstrable competence in the determination of the measurand(s) under test, using validated methods known to be highly precise and accurate, and comparable to methods in general use. The laboratories may, in some situations, be Reference Laboratories.

e) **Consensus values from participant laboratories** — using statistics described in A.1.3 with consideration of the effects of extreme values.

**A.1.2** Assigned values should be determined to evaluate participants fairly, yet to encourage interlaboratory and intermethod agreement. This is accomplished through selection of common comparison groups, wherever possible, and the use of common assigned values.

**A.1.3** The following statistics may be appropriate when assigned values are determined by consensus techniques:

a) **qualitative value** — consensus of a predetermined majority percentage (usually expressed on a nominal or ordinal scale);

b) **quantitative value** — "average" for an appropriate comparison group such as

   i) mean, which may be weighted or transformed (e.g. trimmed or geometric mean),

   ii) median, mode or other robust measure.

**A.1.4** Where appropriate, the uncertainty of assigned values should be determined using procedures described in *Guide to the Expression of Uncertainty in Measurement*.

**A.1.5** Extreme results are treated as follows.

a) When participants' results are used to determine assigned values, techniques should be in place to minimize the influence of extreme results. This can be accomplished with robust statistical methods or by removing outliers prior to calculation (see ISO 5725-2). In larger or routine schemes, it may be possible to have automated outlier screens.

b) If results are removed as outliers, they should be removed only for calculation of summary statistics. These results should still be evaluated within the proficiency scheme and be given the appropriate performance rating.

**A.1.6** Other considerations are as follows.

a) Ideally, if assigned values are determined by reference or participant consensus, the coordinator should have a procedure to establish the trueness of the assigned values and for reviewing the distribution of the data.

b) The coordinator should have criteria for the acceptability of an assigned value in terms of its uncertainty.

## A.2 Calculation of performance statistics

### A.2.1 Performance on single test items

**A.2.1.1** Proficiency test results often need to be transformed into a performance statistic, to aid interpretation and to allow comparison with defined goals. The objective is to measure the deviation from the assigned value in a manner that allows comparison with performance criteria. Techniques may range from no processing required to complex statistical transformations.

**A.2.1.2** Performance measures should be meaningful to scheme participants. Therefore, measures should relate to the application needs for the test and be well understood or traditional within a particular field.

**A.2.1.3** Variability measures are often used for calculation of performance statistics and in summary reports of proficiency testing schemes. Common examples of such variability measures for an appropriate comparison group include:

a) standard deviation (SD)

b) coefficient of variation (CV) or relative standard deviation (RSD)

c) percentiles, median absolute deviation or other robust measures.

**A.2.1.4** For qualitative results, no calculation is usually necessary.

Commonly used statistics for quantitative results are listed below in order of increasing degree of transformation of participants' results.

a) Difference $(x - X)$, where "$x$" is the participant's result and "$X$" is the assigned value.

b) Percent difference,

$$\frac{(x-X)}{X}\times 100$$

c) Percentile or rank.

d) $z$ scores, where

$$z=\frac{x-X}{s}$$

and $s$ is an appropriate estimate/measure of variability which is selected to meet the requirements of the scheme. This model can be used both in the situation where $X$ and $s$ are derived from participants' results or when $X$ and $s$ are not derived from (all) the participant results. [For example, when assigned values and variability are specified; refer to 4.2 of *International Harmonized Protocol for Proficiency Testing of (Chemical) Analytical Laboratories*.]

e) $E_n$ numbers (typically used in measurement comparison schemes), where

$$E_n=\frac{x-X}{\sqrt{U_{\text{lab}}^2+U_{\text{ref}}^2}}$$

and $U_{\text{lab}}$ is the uncertainty of a participant's result and $U_{\text{ref}}$ is the uncertainty of the reference laboratory's assigned value.

**A.2.1.5** Considerations are as follows.

a) The simple difference between the participant's result and the assigned value may be adequate

to determine performance, and is most easily understood by participants. The quantity $(x - X)$ is called the "estimate of laboratory bias" in ISO 5725-4.

b) The percent difference adjusts for concentration, and is well understood by participants.

c) Percentiles or ranks are useful for highly disperse or skewed results, ordinal responses, or when there are a limited number of different responses. This technique should be used with caution (see 6.6.5).

d) Transformed results may be preferred, or necessary, depending on the nature of the test. For example, dilution-based results are a form of geometric scale, transformable by logarithms.

e) If statistical criteria are used (e.g. $z$ scores), the estimates of variability should be reliable; that is, based on enough observations to reduce the influence of extreme results and achieve low uncertainty.

### A.2.2 Combined performance scores

**A.2.2.1** Performance may be evaluated on the basis of more than one result in a single proficiency test round. This occurs when there is more than one test item for a particular measurand, or a family of related measurands. This would be done to provide a more comprehensive evaluation of performance.

Graphical methods such as the Youden Plot or a plot showing Mandel's $h$-statistics are effective techniques for interpreting performance (see ISO 5725-2).

Examples are as follows.

a) Composite score for the same measurand:

- number of satisfactory results;
- average or summed $z$ score;
- average absolute difference (in units or percent);
- summed absolute difference (or square difference);

b) Composite score for different measurands:

- number (or percent) of satisfactory results;
- average absolute $z$ score;
- average absolute difference relative to the evaluation limits.

**A.2.2.2** Considerations are as follows.

a) Scores may be transformed (if necessary) so that they all follow the same assumed distribution (e.g. Gaussian for $z$ scores or chi square for squared differences).

b) There should be a check for extreme values that could heavily influence a quantitative composite score.

## A.3 Evaluation of performance

### A.3.1 Initial performance

Criteria for performance evaluation should be established after taking into account whether the performance measure involves certain features.

**A.3.1.1** These features are the following.

a) **Expert consensus:** where the advisory group, or other qualified experts, directly determine whether reported results are fit for the purpose. Expert consensus is the typical way to assess results for qualitative tests.

b) **Fitness for purpose:** considering, for example, method performance specifications and participants' recognized level of operation.

c) **Statistical determination for scores:** where criteria should be appropriate for each score. Common examples of application of scores are:

i) for $z$ scores:

$|z| \leq 2$ = satisfactory

$2 < |z| < 3$ = questionable

$|z| \geq 3$ = unsatisfactory

ii) for $E_n$ numbers:

$|E_n| \leq 1$ = satisfactory

$|E_n| > 1$ = unsatisfactory

d) **Consensus of participants:** the range of scores or results used by some percentage of participants, or from a reference group, such as:

- central percentage (80 %, 90 % or 95 %) satisfactory, or
- one-sided percentage (lowest 90 %) satisfactory.