INTERNATIONAL STANDARD

ISO 13528

First edition
2005-09-01

Statistical methods for use in proficiency testing by interlaboratory comparisons

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

Reference number
ISO 13528:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

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*.

0 Introduction

0.1 The aims of proficiency testing

Proficiency testing by interlaboratory comparisons is used to determine the performance of individual laboratories for specific tests or measurements, and to monitor the continuing performance of laboratories. The Introduction to ISO/IEC Guide 43-1:1997 should be consulted for a full exposition of the purposes of proficiency testing. In statistical language, the performance of laboratories can be described by three properties: laboratory bias, stability and repeatability. Laboratory bias and repeatability are defined in ISO 3534-1, ISO 3534-2 and ISO 5725-1. The stability of a laboratory's results is measured by intermediate precision as defined in ISO 5725-3.

Laboratory bias may be assessed by tests on reference materials, when these are available, using the procedure described in ISO 5725-4. Otherwise, proficiency testing by interlaboratory comparisons provides a generally available means of obtaining information about laboratory bias, and the use of data from proficiency tests to obtain estimates of laboratory bias is an important aspect of the analysis of such data. However, stability and repeatability will affect data obtained in proficiency tests, so that it is possible for a laboratory to obtain data in a round of a proficiency test which indicate bias that is actually caused by poor stability or poor repeatability. It is therefore important that these aspects of laboratory performance are assessed regularly.

Stability may be assessed by re-testing of retained samples, or by making regular measurements on a reference material or an in-house reference material (a stock of material established by a laboratory to use as private reference material). These techniques are described in ISO 5725-3. Stability may also be assessed by plotting estimates of laboratory bias derived from proficiency tests in control charts. This can provide information about laboratory performance that is not apparent from the examination of the results of individual rounds of proficiency testing schemes, and is another important aspect of the analysis of such data.

Data suitable for assessing repeatability may be generated by tests carried out in the normal course of the work of a laboratory, or by extra tests carried out within a laboratory specifically to assess repeatability. Consequently, the assessment of repeatability is not necessarily an important aspect of proficiency testing, although it is important that laboratories monitor their repeatability in some way. Repeatability may be assessed by plotting ranges of duplicate measurements on a control chart as described in ISO 5725-6.

The flowchart (Figure 1) illustrates how the techniques described in this International Standard are to be applied.

0.2 ISO/IEC Guide 43

ISO/IEC Guide 43-1 describes different types of proficiency testing schemes and gives guidance on the organization and design of proficiency testing schemes. ISO/IEC Guide 43-2 gives guidance on the selection and use of proficiency testing schemes by laboratory accreditation bodies. Those documents should be consulted for detailed information in those areas (the information is not duplicated here). ISO/IEC Guide 43-1 contains an annex that briefly describes the statistical methods that are used in proficiency testing schemes.

This International Standard is complementary to ISO/IEC Guide 43, providing detailed guidance that is lacking in that document on the use of statistical methods in proficiency testing. ISO 13528 is to a large extent based on a harmonized protocol for the proficiency testing of analytical laboratories [1], but is intended for use with all measurement methods.

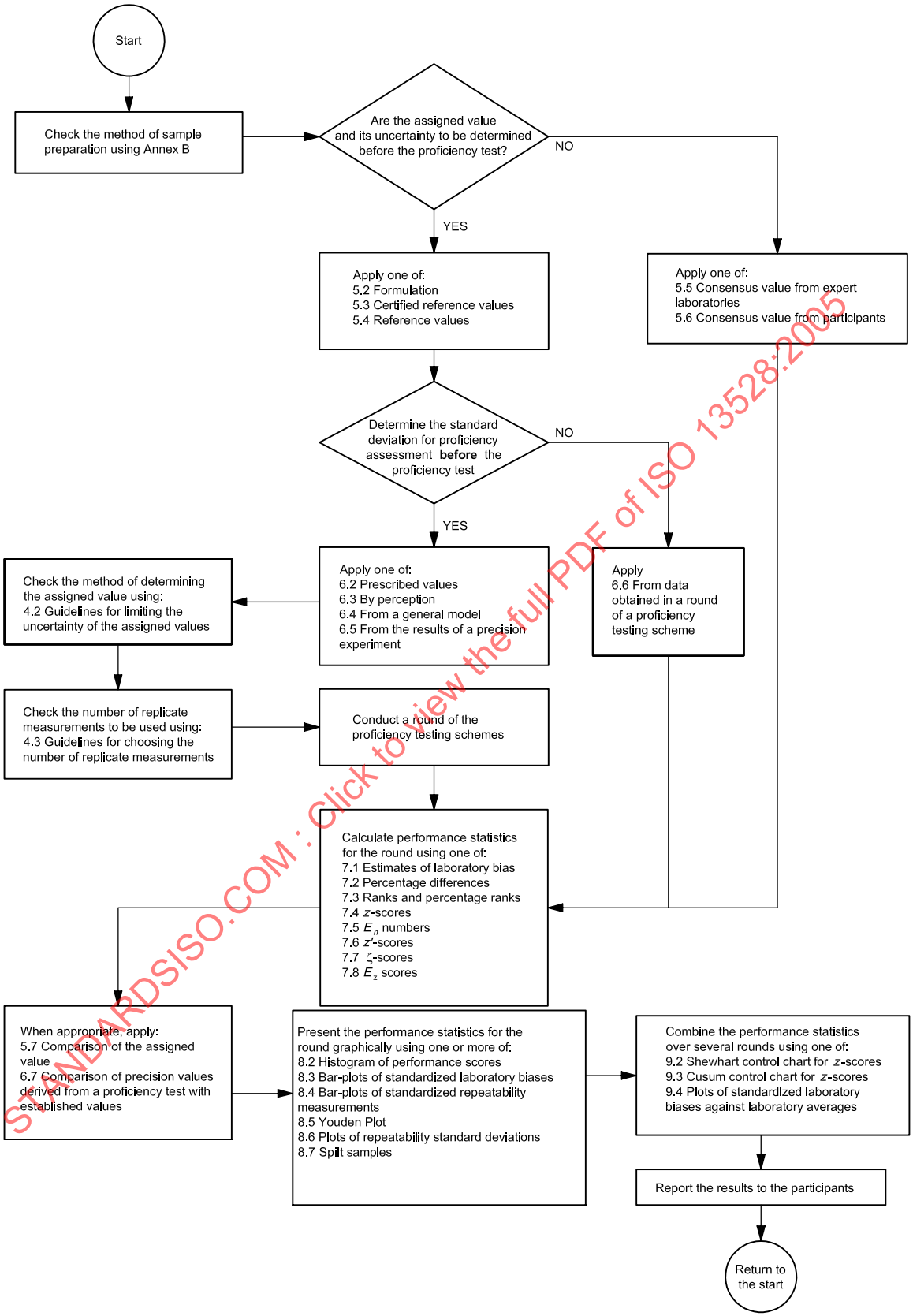


Figure 1 — Flowchart showing the activities requiring the use of statistical methods when operating a proficiency testing scheme

Statistical methods for use in proficiency testing by interlaboratory comparisons

1 Scope

This International Standard complements ISO Guide 43 (all parts) by providing detailed descriptions of sound statistical methods for organizers to use to analyse the data obtained from proficiency testing schemes, and by giving recommendations on their use 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 do not exhibit evidence of an unacceptable level of bias.

It is applicable to quantitative data but not to qualitative data.

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:—[1)], *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 Guide 43-1:1997, *Proficiency testing by interlaboratory comparisons — Part 1: Development and operation of proficiency testing schemes*

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 and the following apply.

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

NOTE Adapted from ISO/IEC Guide 43-1.

3.2
proficiency testing
determination of laboratory testing performance by means of interlaboratory comparisons

1) To be published.

3.3
assigned value
value attributed to a particular quantity and accepted, sometimes by convention, as having an uncertainty appropriate for a given purpose

3.4
standard deviation for proficiency assessment
measure of dispersion used in the assessment of proficiency, based on the available information

3.5
***z*-score**
standardized measure of laboratory bias, calculated using the assigned value and the standard deviation for proficiency assessment

3.6
coordinator
organization (or person) with responsibility for coordinating all of the activities involved in the operation of a proficiency testing scheme

4 Statistical guidelines for the design and interpretation of proficiency tests
(see ISO/IEC Guide 43-1:1997, 5.4.2.)

4.1 Action and warning signals

4.1.1 This International Standard describes some simple numerical or graphical criteria that should be applied to the data obtained in a proficiency test to see if they give rise to action or warning signals. Even in a well-run laboratory, with experienced staff, anomalous results may sometimes be obtained. Also, it is possible that a standardized measurement method, even though it has been validated by a precision experiment, may contain faults that become apparent only after several rounds of a proficiency testing scheme. The proficiency scheme itself may contain faults. For these reasons, the criteria given here shall not be used to condemn laboratories, as being unfit to perform the measurement method under examination. If proficiency testing is used to condemn laboratories, then it shall be necessary to devise appropriate criteria for that purpose.

4.1.2 The criteria given here are designed so that, when the standard deviation for proficiency assessment is based on observed performance (using one of the methods described in 6.4 to 6.6), the criteria give action signals when results are so exceptional as to merit investigation and corrective action.

4.1.3 The coordinator should have an understanding of the major sources of variability that can be anticipated in proficiency test data for the measurement in question. The first step in any analysis should be to examine the distribution of results for evidence of unanticipated sources of variability. For example, a bimodal distribution might be evidence of a mixed population of results caused by different methods, contaminated samples or poorly worded instructions. In this situation, the concern should be resolved before proceeding with analysis or evaluation. Accrediting bodies shall have policies for response to unacceptable performance in proficiency testing. Follow-up actions are determined by that policy or by the laboratory's quality procedures. However, there are generally recommended actions when a laboratory produces an unacceptable result in a proficiency test. Guidance for actions by laboratories in response to unsuccessful performance on a proficiency test is given in 4.1.4.

4.1.4 In schemes where the standard deviation for proficiency assessment is based on observed performance, when a result gives an action signal, the laboratory shall decide what investigations and corrective actions are appropriate, in consultation with the coordinator or an accreditation body if necessary. Unless there is a valid reason not to do so, the laboratory shall examine its procedures and identify one or more corrective actions that, according to staff in the laboratory, are likely to prevent the recurrence of such results. The laboratory may ask the coordinator for advice on possible causes of its problem, or may ask the coordinator to consult other experts. The laboratory shall take part in further rounds of the proficiency testing scheme to assess the effectiveness of the corrective actions. Appropriate corrective actions may be one of the following:

a) checking that staff understand and follow the measurement procedure;

b) checking that all details of the measurement procedure are correct;

c) checking the calibration of equipment and the composition of reagents;

d) replacing suspect equipment or reagents;

e) comparative tests of staff, equipment and/or reagents with another laboratory.

The use of the results of proficiency tests by laboratory accreditation bodies is described in ISO/IEC Guide 43-2:1997, Clause 6.

4.2 Guidelines for limiting the uncertainty of the assigned value

The assigned value X has a standard uncertainty u_X that depends on the method that is used to derive it, and also, when it is derived from tests in several laboratories, on the number of laboratories and, perhaps, on other factors. Methods for calculating the standard uncertainty of the assigned value are given in Clause 5.

The standard deviation for proficiency testing $\hat{\sigma}$ is used to assess the size of estimates of laboratory bias found in a proficiency test. Methods for obtaining the standard deviation for proficiency testing are given in Clause 6 and criteria that compare it with estimates of laboratory bias are given in Clause 7.

If the standard uncertainty u_X of the assigned value is too large in comparison with the standard deviation for proficiency testing $\hat{\sigma}$, then there is a risk that some laboratories will receive action and warning signals because of inaccuracy in the determination of the assigned value, not because of any cause within the laboratories. For this reason, the standard uncertainty of the assigned value shall be established and shall be reported to laboratories participating in proficiency testing schemes (see ISO/IEC Guide 43-1:1997, A.1.4 and A.1.6).

If

$$u_X \leqslant 0{,}3\,\hat{\sigma} \tag{1}$$

then the uncertainty of the assigned value is negligible and need not be included in the interpretation of the results of the proficiency test.

If these guidelines are not met, then the coordinator shall consider the following.

a) Look for a method for determining the assigned value such that its uncertainty meets the above guideline.

b) Use the uncertainty of the assigned value in the interpretation of the results of the proficiency test (see 7.5 on E_n numbers or 7.6 on the z'-score).

c) Inform the participants in the proficiency test that the uncertainty of the assigned value is not negligible.

EXAMPLE Suppose that the assigned value X is determined as the average $\bar{x}$ of the results of tests in 11 laboratories, and that the standard deviation for proficiency testing is determined as the standard deviation s of these same 11 results, so $\hat{\sigma} = s$. As a first approximation, the standard uncertainty of the assigned value in this situation may be estimated by $u_X = s/\sqrt{11} = 0{,}3\,s$, so that the requirement appears to be met. However, the requirement cannot be met in this situation with fewer than 11 laboratories. Further, the uncertainty of the assigned value will be larger than $s/\sqrt{11}$ if the samples suffer from non-homogeneity or instability, or if there is a factor that causes a common bias in the results of the laboratories (e.g. if they all use the same reference standard).

4.3 Guidelines for choosing the number of replicate measurements

Repeatability variation contributes to the variation between the laboratory biases in a proficiency test. If the repeatability variation is too large in comparison with the standard deviation for proficiency testing, then there

is a risk that repeatability variation will cause the results of the proficiency test to be erratic. In this situation, a laboratory could have large bias in one round, but not the next, and they will have difficulty identifying the cause.

For this reason, when it is considered desirable to limit the influence of repeatability variation, the number of replicate measurements n made by each laboratory in a proficiency test shall be chosen so that:

$$\sigma_r / \sqrt{n} \leqslant 0{,}3\,\hat{\sigma} \qquad (2)$$

where σ_r is the repeatability standard deviation that has been established in a previous interlaboratory experiment.

The justification for the factor of 0,3 is that when this criterion is met, the repeatability standard deviation contributes no more than about 10 % of the standard deviation for proficiency testing.

Further, all laboratories shall carry out the same number of replicate measurements. (The methods of analysis given later in this International Standard assume that this requirement is met.) If the requirement inequality of (2) is not met, then the number of replicate measurements shall be increased, or the results of the proficiency test shall be interpreted with caution.

This approach assumes that laboratories have generally similar repeatability. Cases can arise where this is not so. In such cases, for the methods described in this International Standard to be applied, the following device may be used. The coordinator should fix the number of replicate measurements n, using a typical value for the repeatability standard deviation. Then, each laboratory should check that it satisfies Inequality (2) with its own repeatability standard deviation. If it does not, then it should modify its measurement procedure so that it obtains a test result as the average of some number of determinations chosen so that Inequality (2) is satisfied.

4.4 Homogeneity and stability of samples (see ISO/IEC Guide 43-1:1997, 5.6.2 and 5.6.3)

Methods are given in Annex B for checking that the samples to be used in a proficiency test are adequately homogeneous and stable.

When a method of sample preparation is used such that the homogeneity criterion in Annex B is not met, then replicate samples shall be tested by the participants, or the standard deviation for proficiency testing shall include an allowance for the heterogeneity of the samples, as described in Annex B.

4.5 Operationally defined measurement methods

With an operationally defined measurement method, the measurement result is defined by the measurement procedure. For example, the size distribution of a particulate material may be determined using either square-holed or round-holed sieves. There may be no good reason why one type of sieve should be preferred, but unless the type of sieve is specified, laboratories that use different types of sieves may obtain differing results. If a participant uses a different method from that used to establish the assigned value, then their results may show a bias when no fault in execution is present.

If participants are free to choose between operationally defined methods, no valid consensus may be evident amongst them. Two recourses are available to overcome this problem:

a) when a standardized method is in routine use by the participants, it is used to establish the assigned value, and participants are instructed to use it in the proficiency test;

b) a separate value of the assigned value is produced for each method used.

A similar situation arises when the measurand is specified, but not the procedure, and the same choice has to be made.

4.6 Reporting of data (see ISO/IEC Guide 43-1:1997, 6.2.3)

For the purposes of the calculations required in proficiency testing, it is recommended that individual measurement results should not be rounded by more than $\sigma_p / 2$.

Participants shall be asked to report the actual values of their measurement results. Measurement results shall not be truncated (i.e. results shall not be reported in the form "<0,1" or "less than the detection limit"). Likewise, when a negative result is observed, the actual negative value shall be reported (i.e. results shall not be reported as zero even when logically the measurement result cannot be negative). Participants shall be informed that if they report truncated results on a sample, or zeros when results are negative, then all the data for that sample will be excluded from the analysis. If necessary, the form used to report results may contain a box to allow a participant to indicate that a result is below the detection limit.

4.7 Period of validity of the results of proficiency tests

The period of validity of the result obtained by a laboratory in a single round of a proficiency testing scheme is limited to the time that the laboratory performed the test. Thus, if a laboratory achieves a satisfactory result in a single round, the result shall not be used to support a claim that the laboratory obtained reliable data on any other occasion.

A laboratory that operates a quality system and achieves a history of satisfactory results in many rounds of a proficiency testing scheme shall be entitled to use the results as evidence that it is able to obtain consistently reliable data.

5 Determination of the assigned value and its standard uncertainty

5.1 Choice of method of determining the assigned value

Five ways of determining the assigned value X are described in 5.2 to 5.6. The choice between these methods shall be the responsibility of the coordinator following consultation with technical experts as described in ISO/IEC Guide 43-1. The methods described in 5.5 and 5.6 are unlikely to be applicable when the number of laboratories participating in the scheme is small. The methods of calculating the standard uncertainty, u_X, of the assigned value given in this clause will usually be adequate for the applications for which they are used in this International Standard. Alternative methods may be used provided that they have a sound statistical basis and that the method used is described in the documented plan for the scheme.

The determination of the assigned value shall be the responsibility of the coordinator. The assigned value shall not be disclosed to the participants until they have reported their results to the coordinator. The coordinator shall prepare a report giving details of how the assigned value was obtained, the identities of laboratories involved in its determination, and statements of the traceability and measurement uncertainty of the assigned value.

The *Guide to the expression of uncertainty in measurement* gives guidance on the evaluation of measurement uncertainties.

This International Standard recommends the use of robust statistical methods when it is considered that they are the most appropriate methods to use (for example, as in 5.5 and 5.6). Alternatively, procedures that involve the detection and removal of outliers may be used provided that they have a sound statistical basis and the method that is used is reported. Guidance on the use of tests for outliers is given in ISO 5725-2.

5.2 Formulation [see ISO/IEC Guide 43-1:1997, A.1.1, item a)]

5.2.1 General

The test material may be prepared by mixing constituents in specified proportions, or by adding a specified proportion of a constituent to a base material. In this case, the assigned value X is derived by calculation from the masses used.

The approach is especially valuable when individual samples may be prepared in this way, and it is the proportion of the constituents or of the addition that is to be determined: there is then no need to prepare a bulk quantity and ensure that it is homogeneous. However, when formulation gives samples in which the addition is more loosely bonded than in typical materials, or in a different form, it may be preferable to use another approach.

5.2.2 Standard uncertainty u_X of the assigned value

When the assigned value is calculated from the formulation of the test material, the standard uncertainty is estimated by combination of uncertainties using the approach described in the *Guide to the expression of uncertainty in measurement*. For example, in chemical analyses the uncertainties will usually be those associated with gravimetric and volumetric measurements.

The limitation of this method (in chemical analysis) is that care is needed to ensure that:

a) the base material is effectively free from the added constituent, or that the proportion of the added constituent in the base material is accurately known;

b) the constituents are mixed together homogeneously (where this is required);

c) all sources of error are identified (e.g. it is not always realized that glass absorbs mercury compounds, so that the concentration of an aqueous solution of a mercury compound can be altered by its container);

d) there is no interaction between the constituents and the matrix.

5.2.3 Example: Determination of the cement content of hardened concrete

In this case, concrete specimens may be prepared by weighing out quantities of the constituents (cement, aggregates and water) and mixing them together to form each concrete sample. The approach is satisfactory because the accuracy with the specimens can be prepared is far superior to that of the analytical method used to determine the cement content.

5.3 Certified reference values [see ISO/IEC Guide 43-1:1997, A.1.1 item b)]

5.3.1 General

When the material used in a proficiency test is a certified reference material (CRM), its certified reference value is used as the assigned value X.

5.3.2 Standard uncertainty u_X of the assigned value

When a certified reference material is used as the test material, the standard uncertainty of the assigned value is derived from the information on uncertainty provided on the certificate.

The limitation of this approach is that it can be expensive to provide every participant in a proficiency test with a sample of a certified reference material.

5.3.3 Example: Los Angeles value of aggregates

The “Los Angeles value” is a measure of the mechanical strength of aggregates that are used for road construction, and results of the test are measured in “LA units”. In an exercise to certify a reference material, a large number of samples of a particular aggregate were prepared, and some of these samples were used in an interlaboratory experiment involving 28 laboratories, allowing an assigned value of X_{CRM} = 21,62 LA units to be established with a standard uncertainty of $u_{X;\mathrm{CRM}}$ = 0,26 LA units. The remaining samples of this aggregate could be used in proficiency tests.

5.4 Reference values [see ISO/IEC Guide 43-1:1997, A.1.1 item c)]

5.4.1 General

In this approach, samples of the test material that is to be the reference material (RM) are prepared first, ready for distribution to the participants. A number of the samples are then selected at random and tested along with certified reference materials, in one laboratory, using a suitable measurement method, and under repeatability conditions (as defined in ISO 3534-2). The assigned value X_{RM} of the test material is then derived from a calibration against the certified reference values of the CRMs.

5.4.2 Standard uncertainty u_X of the assigned value

When the assigned value of a test material is derived from the results of a series of tests on that material and on CRM, the standard uncertainty of the assigned value is derived from the test results, and the uncertainties of the certified reference values of the CRM. If the test material and the CRM are not similar (in matrix, composition and level of results), then the uncertainty arising from this is also to be included.

This method allows the assigned value to be established in a manner that is traceable to the certified values of the CRMs, with a standard uncertainty that can be calculated, and avoids the cost of distributing the CRM to all the participants. These are good reasons for preferring it to other methods. However, the method assumes that there are no interactions between the materials used and the test conditions.

The example in 5.4.3 illustrates how the required uncertainty may be calculated in the simple case when the assigned value of a test material is established by direct comparison with a single CRM.

5.4.3 Example: Los Angeles value of aggregates

The CRM described in the example in 5.3 may be used to determine the assigned value for an RM which is another, similar, aggregate. This determination requires a series of tests to be carried out, in one laboratory, on samples of the two aggregates, using the same measurement method, and under repeatability conditions. Let

X_{CRM} is the assigned value for the CRM

X_{RM} is the assigned value for the RM

D_i is the difference (RM – CRM) between the average results for the RM and the CRM on the i^{th} samples

$\bar{D}$ is the average of the differences D_i

Then

$$X_{RM} = X_{CRM} + \bar{D} \tag{3}$$

The standard uncertainty of the assigned value of the RM may be calculated as:

$$u_{X;RM} = \sqrt{u_{X;CRM}^2 + u_D^2} \tag{4}$$

Table 1 gives an example of data that might be obtained in such a series of tests, and shows how the standard uncertainty u_D of the difference is calculated.

With these results,

$$X_{RM} = 21{,}62 + 1{,}73 = 23{,}35 \text{ LA units} \tag{5}$$

and

$$u_{X;\mathrm{RM}} = \sqrt{0{,}26^2 + 0{,}24^2} = 0{,}35 \text{ LA units} \quad (6)$$

where 0,26 is the standard uncertainty of the assigned value of the CRM (given in the example in 5.3), and 0,24 is the standard uncertainty of $\bar{D}$.

Table 1 — Calculation of the average difference between a CRM and a RM, and of the standard uncertainty of this difference

Sample	RM		CRM		Difference in average values
	Test 1 LA units	Test 2 LA units	Test 1 LA units	Test 2 LA units	RM – CRM LA units
1	20,5	20,5	19,0	18,0	2,00
2	21,1	20,7	19,8	19,9	1,05
3	21,5	21,5	21,0	21,0	0,50
4	22,3	21,7	21,0	20,8	1,10
5	22,7	22,3	20,5	21,0	1,75
6	23,6	22,4	20,3	20,3	2,70
7	20,9	21,2	21,5	21,8	–0,60
8	21,4	21,5	21,9	21,7	–0,35
9	23,5	23,5	21,0	21,0	2,50
10	22,3	22,9	22,0	21,3	0,95
11	23,5	24,1	20,8	20,6	3,10
12	22,5	23,5	21,0	22,0	1,50
13	22,5	23,5	21,0	21,0	2,00
14	23,4	22,7	22,0	22,0	1,05
15	24,0	24,2	22,1	21,5	2,30
16	24,5	24,4	22,3	22,5	2,05
17	24,8	24,7	22,0	21,9	2,80
18	24,7	25,1	21,9	21,9	3,00
19	24,9	24,4	22,4	22,6	2,15
20	27,2	27,0	24,5	23,7	3,00
Average difference, $\bar{D}$					1,73
Standard deviation					1,07
Standard uncertainty of $\bar{D}$ (standard deviation / $\sqrt{20}$)					0,24

NOTE The data are measurements of the mechanical strength of aggregate, obtained from the Los Angeles (LA) test.

5.5 Consensus values from expert laboratories [see ISO/IEC Guide 43-1:1997, A.1.1 item d)]

5.5.1 General

As with the *reference values* approach (5.4), samples of the test material are prepared first, ready for distribution to the participants. Some of these samples are then selected at random and analysed by a group of expert laboratories. Alternatively, the group of expert laboratories may be participants in a round of a proficiency testing scheme, when the assigned value and its uncertainty will be derived after the round is completed. The assigned value X is calculated as the robust average of the results reported by the group of expert laboratories, calculated using Algorithm A in Annex C.

Other calculation methods may be used in place of Algorithm A, provided that they have a sound statistical basis and the report describes the method that is used.

5.5.2 Standard uncertainty u_X of the assigned value

When each of p expert laboratories reports a measurement x_i on the test material together with an estimate u_i of the standard uncertainty of the measurement, and the assigned value X is calculated as a robust average using Algorithm A, the standard uncertainty of the assigned value X is estimated as:

$$u_X = \frac{1{,}25}{p} \times \sqrt{\sum_{i=1}^{p} u_i^2} \tag{7}$$

When the expert laboratories do not report standard uncertainties, or when the uncertainties are not validated independently (e.g. by a laboratory accreditation body), the standard uncertainty of the assigned value shall be estimated as described in 5.6.

NOTE The factor 1,25 represents the ratio of the standard deviation of the median to the standard deviation of the arithmetic mean, for large samples ($p > 10$) from a normal distribution. For normally distributed data, the standard deviation of a robust average calculated using the algorithm in Annex C is not known, but will fall somewhere between the standard deviation of the arithmetic mean and the standard deviation of the median, so the formula gives a conservative estimate of the standard uncertainty u_X. For $p < 10$, the appropriate factor is less than 1,25, so the formula is then doubly conservative.

The limitations of this approach are that there may be an unknown bias in the results of the group of expert laboratories, and the claimed uncertainties may not be reliable.

5.5.3 Example: Petrographic analysis of aggregates

This approach may be used when samples of aggregates are to be distributed, and the participants are to determine the petrographic composition of the samples. Classification of aggregates requires skill and experience, and there are no reference materials available, so in this case the consensus of a small group of experts may well be the best way to establish assigned values.

5.6 Consensus value from participants [see ISO/IEC Guide 43-1:1997, A.1.1 item e)]

5.6.1 General

With this approach, the assigned value X for the test material used in a round of a proficiency testing scheme is the robust average of the results reported by all the participants in the round, calculated using Algorithm A in Annex C.

Other calculation methods may be used in place of Algorithm A, provided that they have a sound statistical basis and the report describes the method that is used. For example, the calculation in C.1 may be stopped at Equation (53) when the median is obtained, and the median absolute deviation may be used in place of C.2.

This approach may be particularly useful with an operationally defined measurement method, provided that the method is standardized.

5.6.2 Standard uncertainty u_X of the assigned value

When the assigned value is derived as a robust average calculated using Algorithm A, the standard uncertainty of the assigned value X is estimated as:

$$u_X = 1{,}25 \times s^* / \sqrt{p} \tag{8}$$

where s^* is the robust standard deviation of the results calculated using Algorithm A in Annex C. (Here a "result" for a participant is the average of all their measurements on the test material.)

The limitations of this approach are that:

a) there may be no real consensus amongst the participants;

b) the consensus may be biased by the general use of faulty methodology and this bias will not be reflected in the standard uncertainty of the assigned value calculated as described above.

Neither of these conditions is rare in the determination of trace constituents.

5.6.3 Example: Antibody concentrations

Table 2 gives data from a round of a proficiency test in which concentrations of three allergen-specific IgE (immunoglobulin E) antibodies were determined. Figure 2 displays these same data in histograms.

To apply Algorithm A, the data are first sorted into ascending order, initial robust estimates of the average and standard deviation are calculated, and then the iterative method of the algorithm is applied. Table 3 gives these calculations for the results for the allergen specific IgE antibody d1 from Table 2.

The calculations required by Algorithm A may be carried out in a spreadsheet as follows.

a) **Step 1:** Enter the data in a column, in ascending order, as shown in Table 3 for Iteration 0. Calculate their average and standard deviation (10,91 and 3,13 in Table 3). Calculate initial values for the robust average and robust standard deviation (10,85 and 3,53 in Table 3) using the formulae given in C.1.

b) **Step 2:** Copy the data into the next column, as shown in Table 3 for Iteration 1. Use the initial values for the robust average and robust standard deviation to calculate the cut-off values (5,56 and 16,15 in Table 3) using the formulae given in C.1. Replace data outside the cut-off values by the cut-off values (2,18 is replaced by 5,56 and 16,30 is replaced by 16,15). Calculate the new average and standard deviation of the altered data (11,03 and 2,81 in Table 3). According to the formulae given in Annex C, the robust average is the same as this average (11,03) and the robust standard deviation (3,19) is obtained by multiplying the standard deviation by 1,134.

c) **Step 3:** With a spreadsheet, there is now no need to create further columns of data. Instead, change the calculation of the cut-off values at the top of the second column of data so that they use the robust average and robust standard deviation from the bottom of the same column. This will give the cut-off values (6,24 and 15,82) shown in Table 3 under Iteration 2. The calculation can then be completed by continuing to replace data outside the cut-off values by the cut-off values until the iterations converge. When data are replaced, the spreadsheet will automatically update the averages and standard deviations and cut-off values, but the changes in these values will become progressively smaller until they are no longer significant.

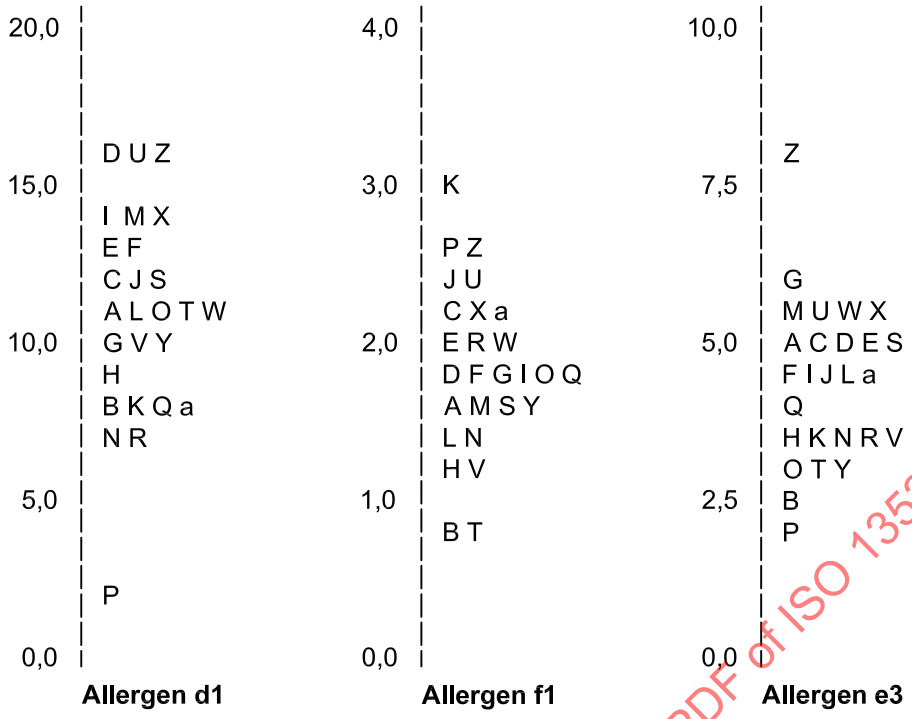
The robust averages and standard deviations for the other two allergen-specific IgE antibodies are calculated similarly.

It is interesting to note that the result for Laboratory P for d1 is not classed as an outlier or a straggler by Grubbs' test. Hence if one relies on the use of outlier tests as described in ISO 5725-2 in this example, the average and standard deviation are calculated from all the data, giving the values shown for *Iteration 0* in Table 3. With the robust method, the low result for Laboratory P and the high results for Laboratories D, U and Z have no influence on the values of the robust estimates. The bar-plots (see Figure 9 in 8.3) would identify the results of some laboratories as being worth investigation. For example, in Figure 9, Laboratory Z can be seen to give large positive z-scores at all three levels.

Table 2 — Concentrations of three allergen specific IgE antibodies (d1, f1 and e3) — Data as reported by p = 27 laboratories

Laboratory	Concentrations		
	d1 kU/l	f1 kU/l	e3 kU/l
A	11,30	1,69	5,02
B	8,29	0,74	2,52
C	11,90	2,23	5,15
D	15,60	1,76	5,15
E	13,40	1,91	4,84
F	12,50	1,71	4,54
G	10,40	1,88	5,94
H	9,38	1,14	3,50
I	14,20	1,74	4,48
J	12,10	2,39	4,75
K	8,10	3,10	3,70
L	10,80	1,39	4,70
M	13,80	1,52	5,59
N	7,00	1,50	3,40
O	10,85	1,80	2,80
P	2,18	2,52	1,88
Q	8,39	1,83	3,80
R	6,95	1,92	3,52
S	11,80	1,58	4,86
T	10,90	0,80	2,80
U	16,30	2,39	5,60
V	9,71	1,21	3,33
W	10,50	1,93	5,35
X	13,60	2,23	5,53
Y	10,10	1,63	3,18
Z	16,07	2,69	8,22
a	8,47	2,16	4,64
p	27	27	27
Robust average x^*	11,03	1,83	4,35
Robust standard deviation s^*	3,04	0,50	1,25

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.



NOTE 1 The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

NOTE 2 The numerical values given in Table 2 are those that will be obtained when the calculation is carried out manually working to two decimal places.

Figure 2 — Concentrations of three allergen specific IgE antibodies — Histograms of data as reported (data from Table 2)

Table 3 — Concentrations of three allergen specific IgE antibodies — Calculation of the robust average and standard deviation for antibody d1

Iteration	0	1	2	3	4	5
$\delta = 1{,}5\ s^*$	—	5,30	4,79	4,62	4,58	4,56
$x^* - \delta$	—	5,56	6,24	6,41	6,45	6,47
$x^* + \delta$	—	16,15	15,82	15,65	15,61	15,59
P	2,18	5,56	6,24	6,41	6,45	6,47
R	6,95	6,95	6,95	6,95	6,95	6,95
N	7,00	7,00	7,00	7,00	7,00	7,00
K	8,10	8,10	8,10	8,10	8,10	8,10
B	8,29	8,29	8,29	8,29	8,29	8,29
Q	8,39	8,39	8,39	8,39	8,39	8,39
a	8,47	8,47	8,47	8,47	8,47	8,47
H	9,38	9,38	9,38	9,38	9,38	9,38
V	9,71	9,71	9,71	9,71	9,71	9,71
Y	10,10	10,10	10,10	10,10	10,10	10,10
G	10,40	10,40	10,40	10,40	10,40	10,40
W	10,50	10,50	10,50	10,50	10,50	10,50
L	10,80	10,80	10,80	10,80	10,80	10,80
O	10,85	10,85	10,85	10,85	10,85	10,85
T	10,90	10,90	10,90	10,90	10,90	10,90
A	11,30	11,30	11,30	11,30	11,30	11,30
S	11,80	11,80	11,80	11,80	11,80	11,80
C	11,90	11,90	11,90	11,90	11,90	11,90
J	12,10	12,10	12,10	12,10	12,10	12,10
F	12,50	12,50	12,50	12,50	12,50	12,50
E	13,40	13,40	13,40	13,40	13,40	13,40
X	13,60	13,60	13,60	13,60	13,60	13,60
M	13,80	13,80	13,80	13,80	13,80	13,80
I	14,20	14,20	14,20	14,20	14,20	14,20
D	15,60	15,60	15,60	15,60	15,60	15,59
Z	16,07	16,07	15,82	15,65	15,61	15,59
U	16,30	16,15	15,82	15,65	15,61	15,59
Average	10,91	11,03	11,03	11,03	11,03	11,03
Standard deviation	3,13	2,81	2,72	2,69	2,68	2,68
New x^*	10,85	11,03	11,03	11,03	11,03	11,03
New s^*	3,53	3,19	3,08	3,05	3,04	3,04

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

5.7 Comparison of the assigned value

When the methods described in 5.2 to 5.4 are used to establish the assigned value X, after each round of a proficiency testing scheme, the robust average x^* derived from the results of the round shall be compared with the assigned value. When the methods described in 5.5 and 5.6 are used to establish the assigned value, the value shall, where possible, be compared with a reference value obtained by a competent laboratory. The standard uncertainty of the difference $x^*–X$ shall be estimated as:

$$\sqrt{\frac{(1{,}25s^*)^2}{p} + u_X{}^2}$$

where

s^* is the robust standard deviation;

p is the number of laboratories.

If the difference is more than twice its uncertainty, the reason shall be sought. Possible reasons are

— bias in the measurement method,

— a common bias in the results of the laboratories,

— failure to appreciate the limitations of the method when using the formulation method described in 5.2 approach,

— bias in the results of the “expert laboratories” when using the “consensus value from expert laboratories" approach, and

— biased participant method(s) or several biased laboratories when the robust consensus mean is used as the assigned value.

5.8 Missing values

When the number of replicate measurements n in a proficiency test is 2 or more, the first step in the analysis of the results will be to calculate the average and standard deviation of each laboratory’s results. The averages are then used, for example, to calculate performance statistics as described in Clause 7 and to prepare histograms or bar-plots as described in 8.2 and 8.3. The standard deviations are used, for example, to prepare the plots of repeatability measures as described in 8.4 and 8.6.

Although all the participants may intend to obtain the same number of replicate measurements, they may not all report this number of measurements, for example, if some tests are spoilt and cannot be repeated. The following procedure is recommended when this happens.

If a laboratory reports at least $0{,}59n$ replicate measurements, then the average and standard deviation of their measurements shall be included in the calculations and treated as if they had reported n measurements. The report shall state how many measurements they reported.

If a laboratory reports less than $0{,}59n$ replicate measurements, then their results shall not be included in the calculation of statistics that affect other laboratories. For example, their results shall not be included in the calculation of the assigned value as described in 5.6, or the calculation of the standard deviation for proficiency assessment as described in 6.6. Their results may be used to calculate their own performance statistics as described in Clause 7, or included in the graphs described in Clause 8, but the report shall state how many measurements they reported and that this was fewer than the number required by the scheme.

NOTE The justification for the multiplier of 0,59 is as follows. The standard deviation of the average of n replicate measurements is $\sigma_r/\sqrt{n}$. If the number of replicate measurements is reduced this standard deviation increases, so if the actual number of replicate measurements is reduced from n to $0{,}59n$, the standard deviation is increased by a factor of 1,3.

This may be considered to be on the borderline of an acceptable increase in the standard deviation. Using the limit of $0{,}59n$ will thus prevent the increase in the standard deviation being greater than this. There is clearly a degree of arbitrariness in the criterion used to derive this rule, so a coordinator may vary it in consultation with the members of the scheme if they wish to.

6 Determining the standard deviation for proficiency assessment
(see ISO/IEC Guide 43-1:1997, A.2.1.3)

6.1 Choice of method

Five approaches to the problem of determining the standard deviation for proficiency assessment $\hat{\sigma}$ are described in 6.2 to 6.6. The choice between these methods shall be the responsibility of the coordinator, in consultation with the members of the scheme and any relevant accreditation bodies, and taking into account any relevant regulations. The method described in 6.6 is unlikely to be applicable when the number of laboratories participating in the scheme is small. The determination of the standard deviation $\hat{\sigma}$ shall be the responsibility of the coordinator. He shall prepare a report giving details of how the standard deviation was obtained.

NOTE ISO/IEC Guide 43-1 uses the symbol s as the standard deviation for proficiency assessment. This is consistent with normal usage when it represents a sample standard deviation. In this International Standard, the standard deviation for proficiency assessment is sometimes derived by other methods, so it is more appropriate to represent it by another symbol. Here $\hat{\sigma}$ has been chosen.

6.2 Prescribed value

6.2.1 General

The standard deviation for proficiency assessment may be set at a value required for a specific task of data interpretation, or it may be derived from a requirement given in legislation.

This approach has the advantage that the standard deviation for proficiency assessment is related directly to a “fitness for purpose” statement for the measurement method.

6.2.2 Example: Aflatoxins in nuts, nut products, dried figs and dried fig products

There is legislation that states that a method used to test for aflatoxins should have a reproducibility coefficient of variation not larger than 50 % when the statutory limit is 10 µg/kg. Thus, if a test material is used in a proficiency testing scheme with a content of aflatoxins of 10 µg/kg, then the legislation implies that the reproducibility standard deviation with this material should be no more than 5 µg/kg. In this case, it would be appropriate to set the standard deviation for proficiency assessment at 5 µg/kg also.

6.3 By perception

6.3.1 General

The standard deviation for proficiency assessment could be set at a value that corresponds to the level of performance that the coordinator and members of the scheme would wish laboratories to be able to achieve.

With this approach, the standard deviation for proficiency assessment becomes equivalent to a “fitness for purpose” statement for the measurement method.

When the standard deviation for proficiency assessment $\hat{\sigma}$ is chosen by prescription or by perception, it is possible that the value chosen is not realistic in relation to the reproducibility of the measurement method. The following method may be used to check that the chosen value of $\hat{\sigma}$ is realistic, provided that information on the repeatability and reproducibility of the method is available. Given

σ_R is the reproducibility standard deviation, and

σ_r is the repeatability standard deviation,

calculate the between-laboratory standard deviation as:

$$\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2} \quad (9)$$

and then calculate the value of the factor ϕ by substituting values of σ_L and σ_r and the chosen value of $\hat{\sigma}$ in Equation (10).

$$\hat{\sigma} = \sqrt{(\phi \times \sigma_L)^2 + (\sigma_r^2 / n)} \quad (10)$$

where n is the number of replicate measurements each laboratory is to perform.

If the value found for ϕ is small (say $\phi < 0{,}5$), it implies that the chosen value of $\hat{\sigma}$ corresponds to a level of reproducibility that laboratories are unable to achieve in practice.

6.3.2 Example 1: Glucose measurement in human serum

Suppose that it is accepted that medical laboratories should be able to determine blood glucose levels within $\pm$ 10 % of the assigned value, although for extremely low concentrations (below 60 mg/dl) a tolerance of $\pm$ 6 mg/dl is acceptable. This information may be used to calculate the standard deviation for proficiency assessment as:

a) for assigned values X below 60 mg/dl:

$\hat{\sigma}$ = 6,0 / 3,0 = 2,0 mg/dl,

b) for assigned values X above 60 mg/dl:

$\hat{\sigma}$ = 0,1 X / 3,0 = 0,033 X mg/dl.

The factor of 3,0 introduced here corresponds to the critical value of 3,0 used in the interpretation of z-scores (see 7.4).

6.3.3 Example 2: Determination of the cement content of hardened concrete

The cement content of concrete is usually measured in terms of the mass in kilograms of cement per cubic metre of concrete (i.e. in kg/m^3). In practice, concrete is produced in grades of quality that have cement contents 25 kg/m^3 apart, and it is desirable that laboratories should be able to identify the grade correctly. For this reason, it is desirable that the chosen value of $\hat{\sigma}$ should be no more that one-half of 25 kg/m^3. A precision experiment produced the following results, for a concrete with an average cement content of 260 kg/m^3: σ_R = 23,2 kg/m^3 and σ_r = 14,3 kg/m^3.

So

$$\sigma_L = \sqrt{23{,}2^2 - 14{,}3^2} = 18{,}3 \text{ kg/m}^3 \quad (11)$$

Taking n as 2, and substituting σ_L = 18,3 kg/m^3, σ_r = 14,3 kg/m^3 and $\hat{\sigma}$ = 12,5 kg/m^3 into Equation (10) gives:

$$12{,}5 = \sqrt{(18{,}3\ \phi)^2 + (14{,}3^2 / 2)} \quad (12)$$

from which one may calculate that ϕ = 0,40. Thus choosing $\hat{\sigma}$ = 12,5 kg/m^3 implies that laboratories are able to achieve a between-laboratory standard deviation lower by a factor of 0,4 than that found in the precision experiment. This is clearly unrealistic.

6.4 From a general model

6.4.1 General

The value of the standard deviation for proficiency testing may be derived from a general model for the reproducibility of the measurement method.

A disadvantage of this approach is that the true reproducibility of a particular measurement method may differ substantially from the value given by the model as the use of a general model implies that the reproducibility depends only on the level of the measurand, and not on the measurand, the measurement procedure, or the sample size.

6.4.2 Example: Horwitz curve

Horwitz [3] gives a general model for the reproducibility of analytical methods that may be used to derive the following expression for the reproducibility standard deviation:

$$\sigma_R = 0{,}02c^{0{,}8495} \tag{13}$$

where c is the concentration of the chemical species to be determined in percent (mass fraction).

6.5 From the results of a precision experiment

6.5.1 General

When the measurement method to be used in the proficiency testing scheme is standardized, and information on the repeatability and reproducibility of the method is available, the standard deviation for proficiency assessment $\hat{\sigma}$ may be calculated using this information, as follows. Given

σ_R is the reproducibility standard deviation, and

σ_r is the repeatability standard deviation,

calculate the between-laboratory standard deviation as:

$$\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2} \tag{14}$$

and then calculate the standard deviation for proficiency assessment as:

$$\hat{\sigma} = \sqrt{\sigma_L^2 + \left(\sigma_r^2 / n\right)} \tag{15}$$

where n is the number of replicate measurements each laboratory is to perform in a round of the scheme.

When the repeatability and reproducibility standard deviations are dependent on the average value of the test results, functional relations will have been derived by the methods described in ISO 5725-2. These relations should then be used to calculate values of the repeatability and reproducibility standard deviations appropriate to the assigned value that is to be used in the proficiency test.

6.5.2 Example: Determination of the cement content of hardened concrete

With the same data used in the example in 6.3, Equation (15) produces a standard deviation for proficiency testing of

$$\hat{\sigma} = \sqrt{18{,}3^2 + \left(14{,}3^2/2\right)} = 20{,}9 \text{ kg/m}^3 \tag{16}$$

assuming that n = 2 replicate measurements are to be made.

6.6 From data obtained in a round of a proficiency testing scheme

6.6.1 General

With this approach, the standard deviation $\hat{\sigma}$ used to assess the proficiency of participants in a round of a scheme is derived from the results reported by the participants in the same round. The standard deviation shall be the robust standard deviation of the results reported by all the participants, calculated using Algorithm A in Annex C. In this context, the result reported by a participant shall be the average of the n replicate measurements obtained by the participant in the round.

Other calculation methods may be used in place of Algorithm A, provided that they have a sound statistical basis and the report describes the method that is used.

A disadvantage of this approach is that the value of $\hat{\sigma}$ may vary substantially from round to round, making it difficult to use values of the z-score for a laboratory to look for trends that persist over several rounds. This disadvantage may be overcome in an established scheme by using a robust pooled value of the standard deviations derived from a number of rounds, calculated using Algorithm S in Annex C.

6.6.2 Example: Antibody concentrations

Tables 2 and 3 provide an example of this approach.

6.7 Comparison of precision values derived from a proficiency test with established values

As a check on the performance of the participants, and to measure the benefit of the scheme to the participants, it is recommended that the coordinator applies the following procedure. The results obtained in each round of a proficiency testing scheme should be used to calculate estimates of the repeatability and reproducibility standard deviations of the measurement method, using the robust methods described in ISO 5725-5. These estimates should be plotted on graphs as time-series, together with values of the repeatability and reproducibility standard deviations obtained in precision experiments (if available).

These graphs should then be examined by the coordinator. If they show that the precision values obtained in the proficiency test differ by a factor of two or more from the values obtained in the precision experiment, then the coordinator should investigate why. If they show that the precision of the measurement method is not improving with time, then it suggests that:

— the participating laboratories are not investigating the causes of action and warning signals, or not implementing corrective actions properly;

— the participating laboratories are not able to identify the causes of action and warning signals;

— the method is in a state of statistical control and reliable conclusions may be based on the data given by the method.

7 Calculation of performance statistics

7.1 Estimates of laboratory bias [see ISO/IEC Guide 43-1:1997, A.2.1.4 item a)]

7.1.1 General

Let x represent the result (or the average of the results) reported by a participant for the measurement of one characteristic of the test material in one round of a proficiency testing scheme.

Then an estimate of the bias D of the laboratory, when measuring that characteristic, may be calculated as:

$$D = x - X \tag{17}$$

where X is the assigned value.

Performance statistics that involve the absolute value $|D|$ of the bias of the laboratory or D^2 should not be used, because they conceal the sign of the bias.

7.1.2 Interpretation of laboratory biases

When a participant reports a result that gives rise to a laboratory bias greater than 3,0 $\hat{\sigma}$ or less than –3,0 $\hat{\sigma}$, then the result shall be considered to give an "action signal". Likewise, a laboratory bias above 2,0 $\hat{\sigma}$ or below –2,0 $\hat{\sigma}$ shall be considered to give a "warning signal". A single "action signal" in one round, or two "warning signals" in successive rounds, shall be taken as evidence that an anomaly has occurred that requires investigation. This criterion is equivalent to that given in 7.4 for z-scores in the sense that it will give the same action and warning signals.

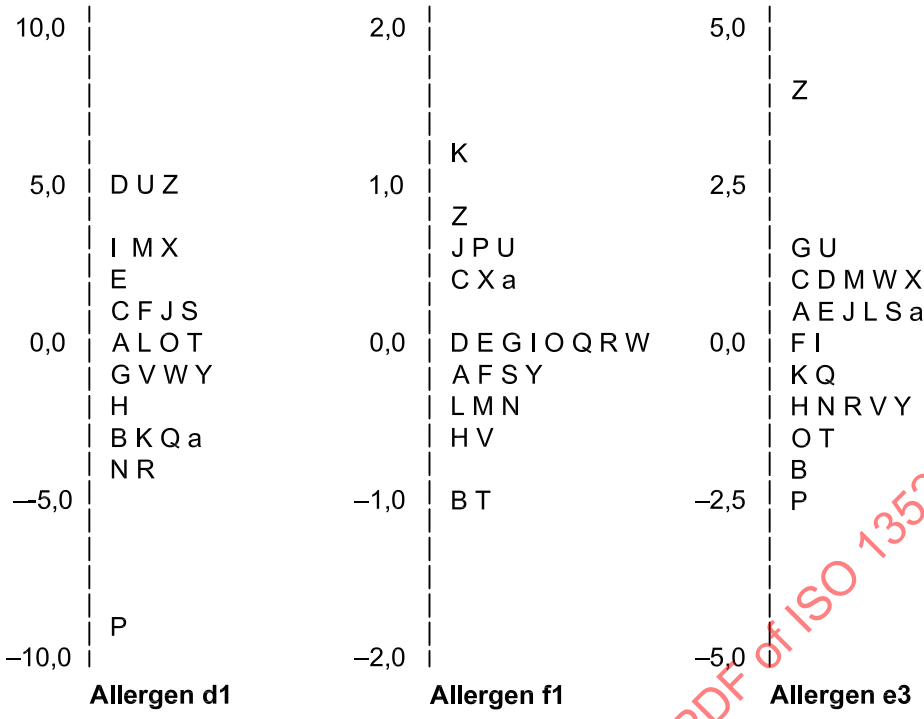
The justification for the use of the multipliers of 2,0 and 3,0 here (and in the other similar criteria given later) is as follows. If X and $\hat{\sigma}$ are good estimates of the mean and standard deviation of the population from which the x values are derived, and the underlying distribution is normal, then the D values will be approximately normally distributed with a mean of zero and a standard deviation $\hat{\sigma}$. Under these circumstances only about 0,3 % of estimated laboratory biases would be expected to fall outside the range $-3{,}0\,\hat{\sigma} < D < 3{,}0\,\hat{\sigma}$, and only about 5 % would be expected to fall outside the range $-2{,}0\,\hat{\sigma} < D < 2{,}0\,\hat{\sigma}$. Because these probabilities are so low, it is unlikely that action signals will occur by chance when no real problem exists, so there is a reasonable chance of identifying the reason for an anomaly when an action signal is given.

When the standard deviation for proficiency assessment is fixed by either of the methods described in 6.2, 6.3 or 6.4, it may differ substantially from the reproducibility standard deviation, and the probabilities of 0,3 % and 5,0 % will then no longer apply.

When the standard deviation for proficiency assessment is fixed by either of the methods described in 6.2 or 6.3, it may be helpful to the participants for the appropriate performance statistic to be used, so that a direct comparison can be made with the prescribed or perceived performance requirement. For example, in the example in 6.3, where the goal for error in glucose samples is ± 10 % of the assigned value, one may follow the example, derive a standard deviation for proficiency assessment of 3,33 %, calculate z-scores and follow 7.4. Equivalently, one may calculate the laboratory bias as a percentage difference (as described in 7.2) and compare it directly with the goal of 10 %.

7.1.3 Example: Antibody concentrations

Table 4 shows the results of applying this method to the data from Table 2, and Figure 3 shows histograms of the estimates of laboratory biases. Comparison of Figures 1 and 2 shows that the laboratory biases have the same distribution as the original data (apart from small effects due to rounding), but they are centred on zero.



NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

Figure 3 — Concentrations of three allergen specific IgE antibodies — Histograms of estimates of laboratory biases (data from Table 4)

Table 4 — Concentrations of three allergen specific IgE antibodies (d1, f1 and e3) — Estimates of laboratory bias

Laboratory	Concentrations		
	d1 kU/l	f1 kU/l	e3 kU/l
A	0,27	–0,14	0,67
B	–2,74	–1,09 W	–1,83
C	0,87	0,40	0,80
D	4,57	-0,07	0,80
E	2,37	0,08	0,49
F	1,47	–0,12	0,19
G	–0,63	0,05	1,59
H	–1,65	–0,69	–0,85
I	3,17	–0,09	0,13
J	1,07	0,56	0,40
K	–2,93	1,27 W	–0,65
L	–0,23	–0,44	0,35
M	2,77	–0,31	1,24
N	–4,03	–0,33	–0,95
O	–0,18	–0,03	–1,55
P	–8,85 W	0,69	–2,47
Q	–2,64	0,00	–0,55
R	–4,08	0,09	–0,83
S	0,77	–0,25	0,51
T	–0,13	–1,03 W	–1,55
U	5,27	0,56	1,25
V	–1,32	–0,62	–1,02
W	–0,53	0,10	1,00
X	2,57	0,40	1,18
Y	–0,93	–0,20	–1,17
Z	5,04	0,86	3,87 A
a	–2,56	0,33	0,29
3,0 $\hat{\sigma}$	9,12	1,50	3,75
2,0 $\hat{\sigma}$	6,08	1,00	2,50
–2,0 $\hat{\sigma}$	–6,08	–1,00	–2,50
–3,0 $\hat{\sigma}$	–9,12	–1,50	–3,75

NOTE 1 When following concentration data, A = Action signal and W = Warning signal.

NOTE 2 The laboratory biases in this table have been derived from the data given in Table 2, using the robust averages in Table 2 as the assigned values for the three levels. The action and warning limits shown at the bottom of the table have been calculated, using the robust standard deviations in Table 2 as the standard deviations for proficiency assessment.

NOTE 3 The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

7.2 Percentage differences [see ISO/IEC Guide 43-1:1997, A.2.1.4 item b)]

7.2.1 General

With the notation as in 7.1, percentage differences are calculated as:

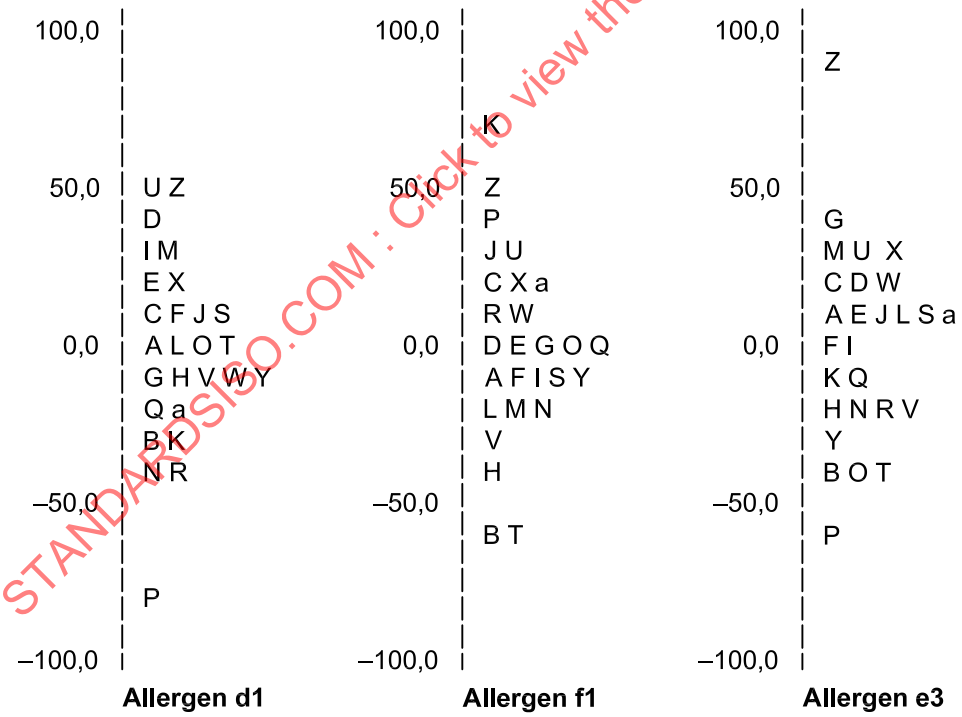
$$D_{\%} = 100\,(x - X)/X \qquad (18)$$

7.2.2 Interpretation of percentage differences

Percentage differences shall be interpreted using equivalent guidelines to those given for laboratory biases in 7.1, i.e. when a participant reports a result that gives rise to a percentage difference above 300 $\hat{\sigma}/X$ % or below $-$ 300 $\hat{\sigma}/X$ %, then the result shall be considered to give an "action signal". Likewise, a percentage difference above 200 $\hat{\sigma}/X$ % or below $-$ 200 $\hat{\sigma}/X$ % shall be considered to give a "warning signal". A single "action signal", or "warning signals" in two successive rounds shall be taken as evidence that an anomaly has occurred that requires investigation.

7.2.3 Example: Antibody concentrations

Table 5 shows the results of applying this method to the data from Table 2, and Figure 4 shows histograms of percentage differences. Comparison of Figures 1 and 3 shows that the percentage differences have the same distribution as the original data (apart from small effects due to rounding), but are centred on zero, just like the estimates of laboratory biases.



NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

Figure 4 — Concentrations of three allergen specific IgE antibodies — Histograms of percentage differences (data from Table 5)

Table 5 — Concentrations of three allergen specific IgE antibodies (d1, f1 and e3) — Percentage differences

Laboratory	Percentage difference		
	d1 %	f1 %	e3 %
A	2	–8	15
B	–25	–60 W	–42
C	8	22	18
D	41	–4	18
E	21	4	11
F	13	–7	4
G	–6	3	37
H	–15	–38	–20
I	29	–5	3
J	10	31	9
K	–27	69 W	–15
L	–2	24	8
M	25	–17	29
N	–37	–18	–22
O	–2	–2	–36
P	–80 W	38	–57
Q	–24	0	–13
R	–37	5	–19
S	7	–14	12
T	–1	–56	–36
U	48	31	29
V	–12	–34	–23
W	–5	5	23
X	23	22	27
Y	–8	–11	–27
Z	46	47	89 A
a	–23	18	7
300 $\hat{\sigma}/X$	82,7	82,0	86,2
200 $\hat{\sigma}/X$	55,1	54,6	57,5
–200 $\hat{\sigma}/X$	–55,1	–54,6	–57,5
–300 $\hat{\sigma}/X$	–82,7	–82,0	–86,2

NOTE 1 When following concentration data, A = Action signal and W = Warning signal.

NOTE 2 The percentage differences in this table have been derived from the data given in Table 2, using the robust averages in Table 2 as the assigned values for the three levels. The action and warning limits shown at the bottom of the table have been calculated using the robust standard deviations in Table 2 as the standard deviations for proficiency assessment.

7.3 Ranks and percentage ranks [see ISO/IEC Guide 43-1:1997, A.2.1.4 item c)]

7.3.1 General

With results from p laboratories in a round of a proficiency test, the ranks are derived by assigning the rank of 1 to the laboratory that reports the lowest result, assigning 2 to the laboratory that reports the next lowest result, and so on, until the laboratory that reports the highest result is assigned the rank of p. If two or more results are equal, they are assigned the same average rank. For example, in Table 2, laboratories C and X both reported a concentration of 2,23 for f1. They share ranks 21 and 22, so in Table 6 they are both assigned a rank of 21,5. If the trial involves several measurands, ranks are assigned for each measurand separately.

If the ranks are denoted by i = 1, 2, ..., p, then the percentage ranks are calculated as $100\,(i - 0{,}5)/p$ %. An example of the calculation of percentage ranks is shown in Table 6.

7.3.2 Interpretation of ranks and percentage ranks

The interpretation of ranks or percentage ranks does not involve an assumption that the data follow a particular probability distribution, and their derivation does not use the assigned value or the standard deviation for proficiency assessment. Hence, ranks and percentage ranks provide a simple method of identifying the laboratories that report the most extreme results. They are of particular use in the early rounds of a proficiency scheme when they can be used to identify the laboratories where improvements in performance are most likely to be achieved. However, the warning given in ISO Guide 43-1:1997, 6.6.5 should be noted: *"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."*

Table 6 — Concentrations of three allergen specific IgE antibodies (d1, f1 and e3) — Ranks and percentage ranks

Laboratory	Rank			Percentage rank		
	d1	f1	e3	d1 %	f1 %	e3 %
A	16	10	19	57	35	69
B	5	1	2	17	2	6
C	18	21,5	20,5	65	78	74
D	25	13	20,5	91	46	74
E	21	17	17	76	61	61
F	20	11	13	72	39	46
G	11	16	26	39	57	94
H	8	3	8	28	9	28
I	24	12	12	87	43	43
J	19	23,5	16	69	85	57
K	4	27	10	13	98	35
L	13	5	15	46	17	54
M	23	7	24	83	24	87
N	3	6	7	9	20	24
O	14	14	3,5	50	50	11
P	1	25	1	2	91	2
Q	6	15	11	20	54	39
R	2	18	9	6	65	31
S	17	8	18	61	28	65
T	15	2	3,5	54	6	11
U	27	23,5	25	98	85	91
V	9	4	6	31	13	20
W	12	19	22	43	69	80
X	22	21,5	23	80	78	83
Y	10	9	5	35	31	17
Z	26	26	27	94	94	98
a	7	20	14	24	72	50

7.4 z-scores [See ISO/IEC Guide 43-1:1997, A.2.1.4 item d)]

7.4.1 General

With notation as in 7.2, the z-score is calculated as:

$$z = \left(x - X\right)/\hat{\sigma} \tag{19}$$

where $\hat{\sigma}$ is the standard deviation for proficiency assessment.

NOTE ISO/IEC Guide 43 uses the symbol s for the standard deviation in the definition of the z-score. This is appropriate when this quantity is derived as the standard deviation of a number of results, but not in other cases (for example, when it is calculated from the results of a precision experiment, or by reference to a general model as in 6.4).

7.4.2 Interpretation of z-scores

When a participant reports a result that gives rise to a z-score above 3,0 or below –3,0, then the result shall be considered to give an "action signal". Likewise, a z-score above 2,0 or below –2,0 shall be considered to give a "warning signal". A single "action signal", or "warning signals" in two successive rounds, shall be taken as evidence that an anomaly has occurred that requires investigation.

With proficiency testing schemes that involve very large numbers of laboratories (e.g. over 100 laboratories), the Normal probability plots as shown in 7.9 and/or Figure 6 may be used to supplement the interpretation of the z-scores. At the other extreme, when there are only a small number of laboratories (e.g. fewer than 10 laboratories), no signal may be given. In this case, the graphical methods that combine scores over several rounds will provide more useful indications of the performance of the laboratories than the results of individual rounds.

7.4.3 Example: Antibody concentrations

The z-scores derived using the robust averages and standard deviations calculated as shown in Table 2 are given in Table 7, and as histograms in Figure 5. Comparison of Figures 1 and 4 shows that the z-scores have the same distribution as the original data (apart from small effects due to rounding), but are centred on zero, just like the estimates of laboratory biases.

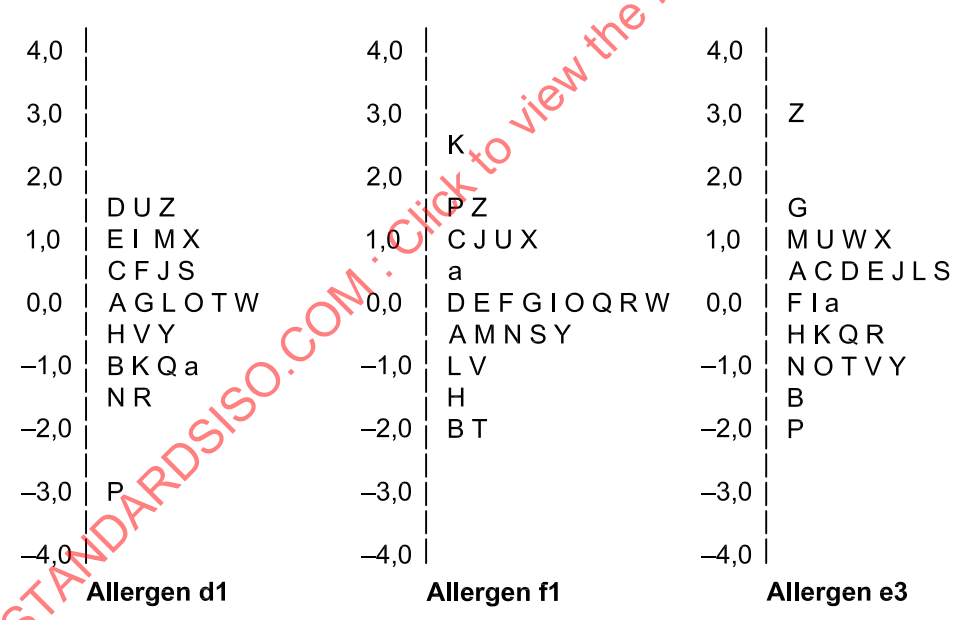


Figure 5 — Concentrations of three allergen specific IgE antibodies — histograms of z-scores (data from Table 7)

Table 7 — Concentrations of three allergen specific IgE antibodies (d1, f1 and e3) — z-scores

Laboratory	z-score		
	d1	f1	e3
A	0,09	–0,28	0,54
B	–0,90	–2,18 W	–1,46
C	0,29	0,80	0,64
D	1,50	–0,14	0,64
E	0,78	0,16	0,39
F	0,48	–0,24	0,15
G	–0,21	0,10	1,27
H	–0,54	–1,38	–0,68
I	1,04	–0,18	0,10
J	0,35	1,12	0,32
K	–0,96	2,54 W	–0,52
L	–0,08	–0,88	0,28
M	0,91	–0,62	0,99
N	–1,33	–0,66	–0,76
O	–0,06	–0,06	–1,24
P	–2,91 W	1,38	–1,98
Q	–0,87	0,00	–0,44
R	–1,34	0,18	–0,66
S	0,25	–0,50	0,41
T	–0,04	–2,06 W	–1,24
U	1,73	1,12	1,00
V	–0,43	–1,24	–0,82
W	–0,17	0,20	0,80
X	0,85	0,80	0,94
Y	–0,31	–0,40	–0,94
Z	1,66	1,72	3,10 A
a	–0,84	0,66	0,23

NOTE 1 When following concentration data, A = Action signal and W = Warning signal.

NOTE 2 The z-scores in this table have been derived from the data given in Table 2, using the robust averages in Table 2 as the assigned values for the three levels, and using the robust standard deviations in Table 2 as the standard deviations for proficiency assessment. The formula for the z-score in this example is thus $z = (x - X)/\hat{\sigma} = (x - x^*)/s^*$.

7.5 E_n numbers [See ISO/IEC Guide 43-1:1997, A.2.1.4 item e)]

This performance statistic is calculated as:

$$E_n = \frac{x - X}{\sqrt{U_{\text{lab}}^2 + U_{\text{ref}}^2}} \tag{20}$$

where

X is the assigned value determined in a reference laboratory;

U_{ref} is the expanded uncertainty of X;

U_{lab} is the expanded uncertainty of a participant's result x.

In contrast to the critical values of 2,0 and 3,0 used with z-scores, it is common to use a critical value of 1,0 with E_n numbers. This is because E_n numbers are calculated using expanded uncertainties in the denominator instead of standard deviations.

NOTE 1 E_n numbers should be used with caution when participants may have a poor understanding of their uncertainty and may not be reporting it in a uniform way. However, incorporating information on uncertainty into the interpretation of results of proficiency tests can play a major role in improving their understanding of this difficult subject.

When the expanded uncertainties are calculated using a coverage factor of 2,0, a critical value of 1,0 for an E_n number is equivalent to the critical value of 2,0 used with z-scores.

NOTE 2 When uncertainties are estimated in a way consistent with the *Guide to the expression of uncertainty in measurement* (GUM), E_n numbers express the validity of the expanded uncertainty estimate associated with each result. A value of $|E_n| < 1$ provides objective evidence that the estimate of uncertainty is consistent with the definition of expanded uncertainty given in the GUM.

7.6 z'-scores

7.6.1 General

With notation as in 7.4, the z'-score is calculated as:

$$z' = (x - X) / \sqrt{\hat{\sigma}^2 + u_X^2} \tag{21}$$

where u_X is the standard uncertainty of the assigned value X.

Equation (21) may be used when the assigned value is not calculated using the results reported by the participants. Thus, it may be used when the assigned value is obtained by the methods described in 5.2, 5.3 and 5.4, and when the method described in 5.5 is used and the expert laboratories do not take part in the proficiency test. When the method described in 5.6 is used, the assigned value is correlated with the results reported by the participants so the use of z'-scores as defined by Equation (21) is not valid.

7.6.2 Interpretation of z'-scores

z'-scores shall be interpreted in the same way as z-scores (see 7.4) and using the same critical values of 2,0 and 3,0.

NOTE The criteria given in 7.1 and 7.2 for the interpretation of laboratory biases and of percentage differences can be modified similarly by replacing $\hat{\sigma}$ by $\sqrt{\hat{\sigma}^2 + u_X^2}$.

7.6.3 Use of z'-scores

Comparison of the formulae for the z-score and the z'-score in 7.4 and 7.6 shows that the z'-scores for a round of a proficiency testing scheme will all be smaller than the corresponding z-scores by a constant factor of

$$\hat{\sigma} / \sqrt{\hat{\sigma}^2 + u_X^2}$$

When the guideline for limiting the uncertainty of the assigned value in 4.2 is met, this factor will fall in the range:

$$0{,}96 \leqslant \hat{\sigma} / \sqrt{\hat{\sigma}^2 + u_X^2} \leqslant 1{,}00 \qquad (22)$$

Thus, in this case, the z'-scores will be nearly identical to the z-scores, and it may be concluded that the uncertainty of the assigned value is negligible.

When the guideline in 4.2 is not met, the difference in magnitude of the z'-scores and z-scores may be such that some z-scores exceed the critical values of 2,0 or 3,0 and so give "warning signals" by an "action signals", whereas the corresponding z'-scores do not exceed these critical values and so do not give signals.

When deciding whether to use either z-scores or z'-scores, the coordinator shall consider the following aspects.

a) Does the uncertainty of the assigned value meet the guidelines in 4.2? If it does, then it is unlikely that there will be any benefit from the use of z'-scores.

b) When the guideline is not met, it is recommended to use z'-scores in spite of their additional complexity and the difficulties of explaining them to users.

c) How severe are the consequences to laboratories when their results give rise to warning or action signals? Are the results used to disqualify laboratories from carrying out the measurement method for some group of users?

7.7 Zeta-scores (ζ)

7.7.1 General

With notation as in 7.4, the ζ-scores is calculated as:

$$\zeta = (x - X) / \sqrt{u_x^2 + u_X^2} \qquad (23)$$

Where u_x is the laboratory's own estimate of the standard uncertainty of its result x, and u_X is the standard uncertainty of the assigned value X.

Equation (23) may be used when the assigned value is not calculated using the results reported by the participants. Thus, it may be used when the assigned value is obtained by the methods described in 5.2, 5.3 and 5.4, and when the method described in 5.5 is used and the expert laboratories do not take part in the proficiency test. When the method described in 5.6 is used, the assigned value is correlated with the results reported by the participants so the use of ζ-scores as defined by Equation (23) is not valid.

NOTE 1 ζ-scores differ from E_n numbers by using standard uncertainties $u_{(x)}$, rather than expanded uncertainties $U_{(x)}$.

NOTE 2 At the present time, it is not common practice to incorporate information provided by participating laboratories on the uncertainty of their measurements in the scores used in proficiency testing schemes. However, such information may become more widely reported. 7.7 is included to provide coordinators with guidance on how such information could be incorporated should it become available. Information on the uncertainty of measurement is now required by ISO/IEC 17025, so for proficiency testing schemes involving laboratories that claim compliance with that International Standard, coordinators need guidance on how such information should be dealt with.

7.7.2 Interpretation of ζ-scores

When there is an effective system in operation for validating laboratories' own estimates of the standard uncertainties of their results, ζ-scores may be used instead of z-scores, and shall be interpreted in the same way as z-scores (see 7.4), using the same critical values of 2,0 and 3,0.

When no such system is in operation, ζ-scores shall be used only in conjunction with z-scores, as an aid for improving the performance of laboratories, as follows. If a laboratory obtains z-scores that repeatedly exceed the critical value of 3,0, they may find it of value to examine their test procedure step by step and derive an uncertainty budget for that procedure. The uncertainty budget will identify the steps in the procedure where the largest uncertainties arise, so that the laboratory can see where to expend effort to achieve an improvement. If their ζ-scores also repeatedly exceed the critical value of 3,0, it implies that their uncertainty budget does not include all significant sources of uncertainty (i.e. they are missing something important).

If a laboratory has a large bias and its uncertainty interval $X \pm U_x$ does not include the assigned value, then it will also have a large ζ-score or E_n number.

7.8 E_z score

The E_z score may be defined as: $E_{z-} = \dfrac{x-(X-U_X)}{U_x}$ and $E_{z+} = \dfrac{x-(X+U_X)}{U_x}$

Here X is the assigned value derived using 7.4 or 7.6, or a reference value derived using 7.5, and U_X is the expanded uncertainty of X.

Also x is a laboratory's value and U_x is the expanded uncertainty of x. U_x can be $2\hat{\sigma}$ or $3\hat{\sigma}$ in 7.4 and 7.6 and U_{lab} in 7.5.

It is common to compare values of E_z with a critical value of 1,0:

a) when both E_{z-} and E_{z+} fall within the range –1,0 to 1,0, the laboratory's performance is satisfactory;

b) when one of E_{z-} and E_{z+} falls outside the range –1,0 to 1,0, the laboratory's performance is questionable;

a) when E_{z-} and E_{z+} are both lower than –1,0 or both higher than 1,0, the laboratory's performance is unsatisfactory.

7.9 An example of the analysis of data when uncertainties are reported

7.9.1 General

Graphs such as those shown in Figures 6 and 7 provide helpful summaries of laboratories' results and their uncertainties. Laboratories that have the largest laboratory biases will also have the largest z-scores. (They will appear at the either end of the graph.)

Table 8 gives authentic data obtained in an exercise in which 181 laboratories reported results with uncertainties for lead in water. The data were reported as mol/l. They have been multiplied by 10^{10} to give more easily managed values. This means that results do not have either the same uncertainty or the same bias. Each result has its own bias D_i resulting from the sum of *Method bias* and *Laboratory bias*; moreover, each method has of course its own Reproducibility standard deviation σ_R. The laboratories used various methods for the determinations, and calculated the uncertainties themselves. The uncertainties are treated here as expanded uncertainties. The data are given in Table 8 as reported, except that they have been sorted so that the results are in ascending order, and the laboratories are numbered in this order. $U = 0$ may indicate a failure to report uncertainty. There are several values in the table of doubtful validity that would merit further investigation in practice, and it is not possible to show the most extreme results on the graphs. In particular, the negative values are not included in the graphs although they are used in the calculations. These data provide an example of where negative results have been reported as required by 4.6, even though negative lead contents are not logically possible.

7.9.2 The assigned value and its uncertainty

The assigned value is calculated as described in 5.6 as the robust average of the results, using Algorithm A in Annex C. This gives an assigned value

$$X = x^* = 605 \times 10^{-10} \text{ mol/l} \tag{24}$$

and a robust standard deviation

$$s^* = 142 \times 10^{-10} \text{ mol/l} \tag{25}$$

According to 5.6, the standard uncertainty of this assigned value is

$$u_X = 1{,}23 \times s^* / \sqrt{181} = 13 \times 10^{-10} \text{ mol/l} \tag{26}$$

7.9.3 The standard deviation for proficiency assessment

The standard deviation for proficiency assessment is obtained as described in 6.6 as the robust standard deviation

$$\hat{\sigma} = s^* = 142 \times 10^{-10} \text{ mol/l} \tag{27}$$

7.9.4 Guidelines for interpreting the uncertainty of the assigned value

According to the guidelines in 4.2, the uncertainty of the assigned value is negligible if

$$u_X \leqslant 0{,}3\hat{\sigma} \tag{28}$$

With $u_X = 1{,}25 \times s^* / \sqrt{p}$ and $\hat{\sigma} = s^*$ as in this example, it can be calculated that Equation (28) is satisfied with $p > 16$. With $p = 181$ laboratories participating, the criterion is easily satisfied. There is therefore no point with these data in considering z'-scores as described in 7.6.

7.9.5 Analysing data for a large number of laboratories using Normal probability plots

Figure 6 shows the results for the 181 laboratories plotted against their percentage ranks (calculated as described in 7.3), using a Normal probability scale for the percentage ranks. Results below 0 mol/l or above 1600×10^{-10} mol/l are not included in this figure.

The z-scores may be calculated as $z = (x - 605)/142$. When a laboratory achieves a z-score above 3,0 or below –3,0, the value of the z-score is shown on this figure alongside the corresponding value.

The cumulative distribution function for the Normal distribution with mean 605×10^{-10} mol/l and standard deviation 142×10^{-10} mol/l is also shown on the figure as the dashed straight line.

The final "cut-off" values used in the robust algorithm are:

$$x^* - 1{,}5s^* = 605 - 1{,}5 \times 142 = 392 \times 10^{-10} \text{ mol/l} \tag{29}$$

and

$$x^* + 1{,}5s^* = 605 + 1{,}5 \times 142 = 818 \times 10^{-10} \text{ mol/l} \tag{30}$$

It can be seen on Figure 6 that the points veer away from the dashed line outside this range. The implication of this is that all the results cannot be considered to be drawn from the same normal population. Points away from the dashed line are drawn from a population with a larger variance than points close to the line.

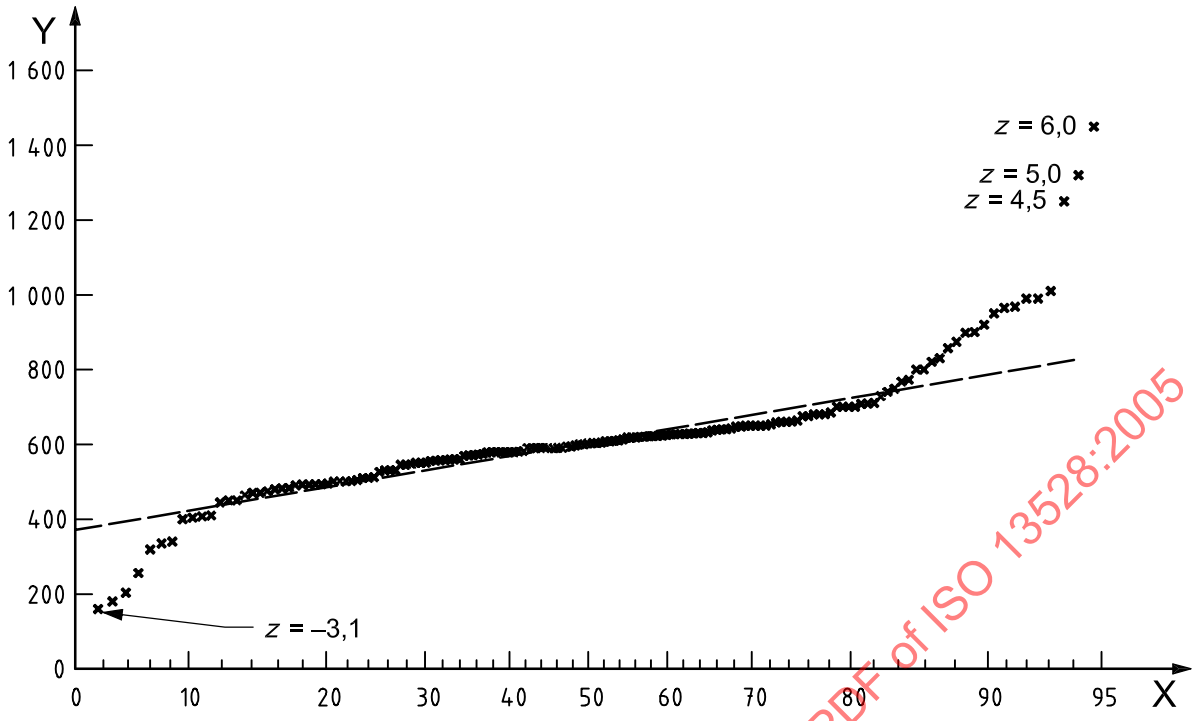
It can also be seen in the figure that results that give z-scores above 3,0, or below −3,0, are well away from the dashed line. This supports a conclusion to treat these z-scores as giving rise to "action" signals. (It is possible with a large number of results that z-scores above 3,0 or below −3,0 are obtained, but when plotted as in Figure 6 all the points lie close to the dashed line. In such a case, the figure would not support a conclusion to treat the z-scores as giving rise to action signals.)

Figure 7 shows the results for just those laboratories that give z-scores within the range ± 3,0, together with vertical lines that indicate the uncertainties that they reported. (Results for two laboratories, numbers 24 and 112, that reported very high uncertainties, have not been included in the figure.) The continuous, horizontal line on this figure represents the assigned value, and the dashed lines on either side of it represent the expanded uncertainty of the assigned value.

If the laboratories carried out valid calculations of the expanded uncertainties of their results, then nearly all of the vertical lines in Figure 7 would cut the region defined by the assigned value ± its expanded uncertainty. However, it can be seen in the figure that there are many laboratories, on both sides of the assigned value, whose vertical lines do not reach this region. This implies that many laboratories have not carried out valid uncertainty calculations; very probably they have not included all the important sources of uncertainty in their calculations.

Table 8 — Determinations of the lead content of water (mol/l × 10^{10}) **by 181 laboratories, together with values for the expanded uncertainties** (U) **of the results as reported by the laboratories**

Lab	Result	U	Lab	Result	U	Lab	Result	U	Lab	Result	U
1	–960 000	0	51	545	43	101	618	224	151	740	20
2	–12 100	0	52	545	123	102	618	170	152	748	3
3	–4 800	0	53	550	8	103	620	25	153	767	113
4	–3 860	0	54	550	55	104	620	40	154	772	213
5	–1 500	0	55	550	5	105	621	6	155	800	60
6	–1 010	0	56	555	79	106	622	9	156	800	150
7	–1 000	0	57	556	30	107	622	6	157	821	203
8	–1 000	0	58	557	28	108	623	18	158	830	10
9	–965	0	59	557	28	109	625	15	159	857	27
10	–483	0	60	559	26	110	626	5	160	874	200
11	160	20	61	560	7	111	627	0	161	898	59
12	180	20	62	560	60	112	627	1 010	162	900	100
13	203	0	63	569	116	113	627	15	163	920	140
14	256	13	64	570	86	114	628	3	164	950	110
15	319	0	65	571	16	115	629	26	165	965	0
16	335	18	66	572	40	116	630	40	166	968	0
17	340	180	67	574	35	117	630	580	167	990	0
18	400	20	68	578	0	118	632	50	168	990	80
19	404	36	69	579	52	119	637	96	169	1 010	0
20	407	0	70	579	35	120	639	83	170	1 250	140
21	410	97	71	579	8	121	640	130	171	1 320	410
22	444	58	72	579	10	122	640	77	172	1 450	460
23	450	20	73	579	17	123	642	20	173	1 640	241
24	450	3 400 000	74	579	87	124	647	63	174	1 900	46
25	463	19	75	580	150	125	647	0	175	2 413	20
26	470	10	76	582	122	126	650	160	176	2 460	0
27	470	30	77	589	57	127	650	30	177	2 900	900
28	474	0	78	589	10	128	650	80	178	10 000	0
29	480	100	79	590	0	129	650	48	179	386 000	31 000
30	482	122	80	590	0	130	650	30	180	670 000	60 000
31	483	241	81	590	45	131	653	5	181	630×10^6	60×10^6
32	490	60	82	590	60	132	658	27			
33	492	25	83	590	0	133	660	20			
34	492	1	84	591	112	134	660	120			
35	493	24	85	591	9	135	660	34			
36	493	5	86	594	4	136	663	32			
37	495	0	87	594	119	137	675	280			
38	500	70	88	597	9	138	675	0			
39	500	10	89	600	20	139	680	8			
40	500	10	90	600	300	140	680	50			
41	501	75	91	603	60	141	680	70			
42	504	0	92	603	24	142	685	0			
43	510	130	93	603	13	143	700	0			
44	510	110	94	604	18	144	700	100			
45	512	6	95	608	30	145	700	110			
46	526	26	96	608	8	146	700	300			
47	530	9	97	609	9	147	708	44			
48	530	40	98	610	61	148	709	48			
49	530	60	99	613	22	149	710	100			
50	545	30	100	618	7	150	729	41			



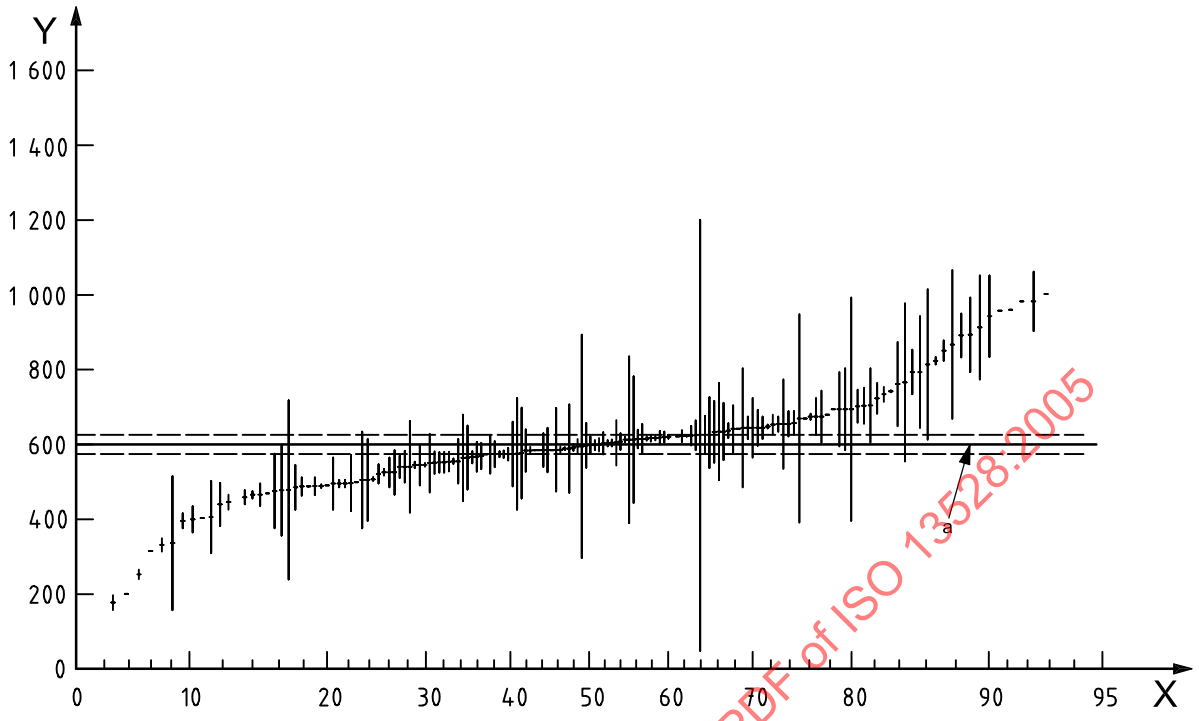
Key

X laboratory-percentage rank, %

Y lead content, mol/l × 10^{10}

NOTE The results for 19 laboratories are not included.

Figure 6 — Normal probability plot of results of determinations of the lead content of water by 162 laboratories



Key

X laboratory-percentage rank, %

Y lead content, mol/l $\times 10^{10}$

[a] Assigned value ± its expanded uncertainty.

NOTE The results for 25 laboratories are not included.

Figure 7 — Normal probability plot of expanded uncertainties for determinations of the lead content of water by 156 laboratories

7.10 Combined performance scores

It is common, within a single round of a proficiency testing scheme, for results to be obtained for more than one test item or for several measurands. In this situation, the results for each test item and for each measurand shall be interpreted as described in 7.2 to 7.9, i.e. the results for each test item and each measurand shall be analysed separately. There are applications when two or more materials with specially designed levels are included in a scheme to measure other aspects of performance, such as repeatability or linearity. In such instances, the coordinator shall provide participants with complete descriptions of the statistical design and procedures that are used. If two similar materials have been incorporated in the scheme with the intention of treating them as a Youden Pair, the special methods described in 8.5 shall be applied.

Further, it is recommended that the graphical methods described in Clause 8 should also be used when results are obtained for more than one test item or for several measurands. They combine scores in ways that do not conceal high values of individual scores, and they may reveal additional information on the performance of laboratories, such as correlation between results for different measurands, not apparent in tables of the individual scores.

In schemes that involve a large number of measurands, a count of the numbers of action and warning signals may be used to allow the laboratories that obtain one or more such signals to be identified. They can be provided with a report containing detailed results using the methods described in 7.2 to 7.9. Laboratories that obtain no signal can be provided with only a brief report.

NOTE The use of composite scores when there is more than one test item for the same measurand (the average or summed z-score, or the average or summed absolute difference or squared difference), or the use of composite scores when there are results for more than one measurand (the average absolute z-score, or the average absolute difference relative to the evaluation limits) is not recommended. The average (or summed) z-score has the serious failing that a high score on one test item can be concealed if other scores are low or if another score is also high but has the opposite sign. The average (or summed) absolute difference and the sum of squared differences also have the serious failing that a high score on one test item can be concealed if other scores are low. The average absolute z-score and the average absolute difference relative to the evaluation limits also have this same failing.

8 Graphical methods for combining performance scores for several measurands from one round of a proficiency test (see ISO/IEC Guide 43-1:1997, A.2.2.1)

8.1 Application

The coordinator shall consider using the performance scores obtained in each round of a proficiency testing scheme to prepare graphs, as described in 8.2 and 8.3. The use of z-scores in these graphs has the advantage that they can be drawn using standardized axes, thereby simplifying their presentation and interpretation. Graphs shall be made available to the participants, enabling each participant to see where his own results fall in relation to those obtained by the other participants. Letter codes or number codes shall be used to represent the participants so that each participant is able to identify his own results but not able to determine which participant obtained any other result. The graphs shall also be made available to the coordinator, enabling them to judge the overall effectiveness of the scheme, and to see if there is a need for reviewing the criterion used to assess proficiency.

8.2 Histograms of performance scores

8.2.1 General

To prepare a histogram of z-scores, collect the z-scores for the measurement of a characteristic from one round of a proficiency testing scheme into a histogram as shown in Figure 8. Use an interval in the histogram of about 0,3 to 0,5 so that the histogram gives a good visual impression. Draw lines at $\pm$ 2,0 and $\pm$ 3,0 to represent the proficiency assessment criterion. Use a range for the histogram of about $\pm$ 6,0. If the results do not span this range, or if the central points are tightly clustered, the analyst may consider an alternative range.

When histograms of laboratory biases or percentage differences are preferred, the equivalent histogram intervals and warning and action limits are as below. In these cases, it may be simpler to derive the action limits directly from the prescribed or perceived requirements instead of calculating them from the standard deviation for proficiency assessment.

Table 9 — Warning and action limits for performance scores

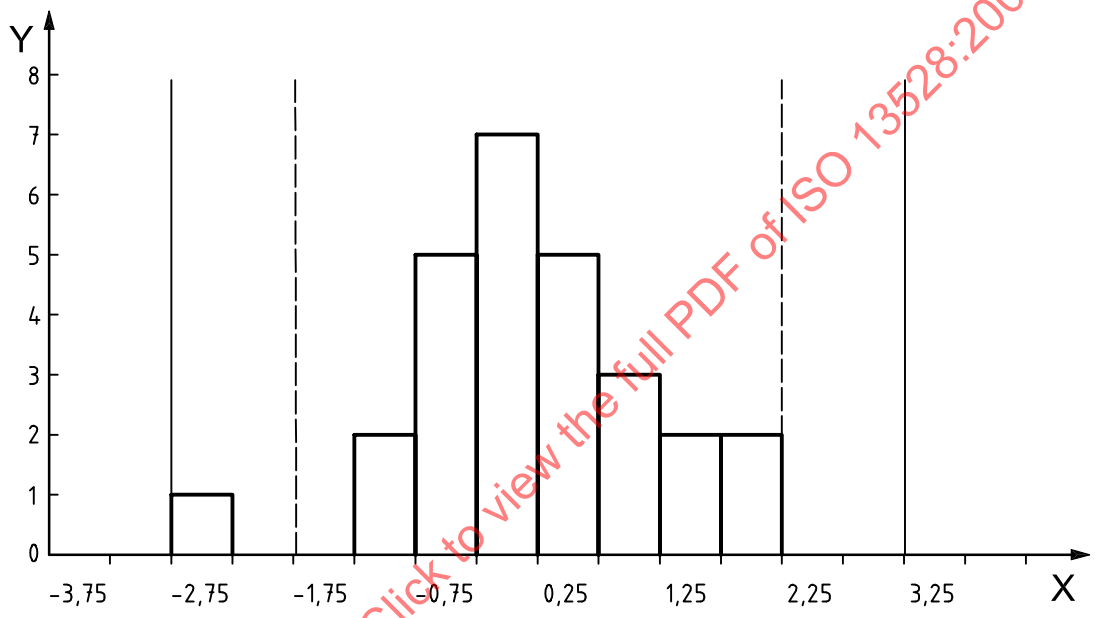
Performance statistic	Histogram interval	Warning limits	Action limits	Histogram range
Laboratory bias	0,3 $\hat{\sigma}$ to 0,5 $\hat{\sigma}$	$\pm$ 2,0 $\hat{\sigma}$	$\pm$ 3,0 $\hat{\sigma}$	$\pm$ 6,0 $\hat{\sigma}$
Percentage difference	30 $\hat{\sigma}$ / X to 50 $\hat{\sigma}$ / X	$\pm$ 200 $\hat{\sigma}$ / X	$\pm$ 300 $\hat{\sigma}$ / X	$\pm$ 600 $\hat{\sigma}$ / X
z-score	0,3 to 0,5	$\pm$ 2,0	$\pm$ 3,0	$\pm$ 6,0

Histograms are a suitable method of graphical presentation when the number of characteristics measured is small, or when a number of dissimilar characteristics are measured. Individual participants can identify the position of their own scores and assess their performance and the need to investigate their methods. A participant who has achieved a high z-score is able to use a histogram to see how exceptional his score is, in comparison with the scores achieved by the other participants.

The coordinator is able to use a histogram to see how frequently participants fail to satisfy the proficiency assessment criterion. If the tails of the histogram extend outside the ± 3,0 limits, then the fault would appear to lie with the measurement method (or methods) being used rather than with individual participants. The measurement method (or methods) should be improved, or the proficiency assessment criterion should be relaxed (by increasing $\hat{\sigma}$). If the histogram lies within the ± 2,0 limits, with perhaps one or two isolated z-scores outside these limits, then it suggests that the proficiency assessment criterion could be made more strict (by reducing $\hat{\sigma}$).

8.2.2 Example: Antibody concentrations

The z-scores for d1 are shown in Figure 8 in the form of a histogram.



Key

X z-score for allergen d1

Y number of laboratories

Figure 8 — Histogram of z-scores for one round of a proficiency test
(data for allergen d1 from Table 7)

8.3 Bar-plots of standardized laboratory biases

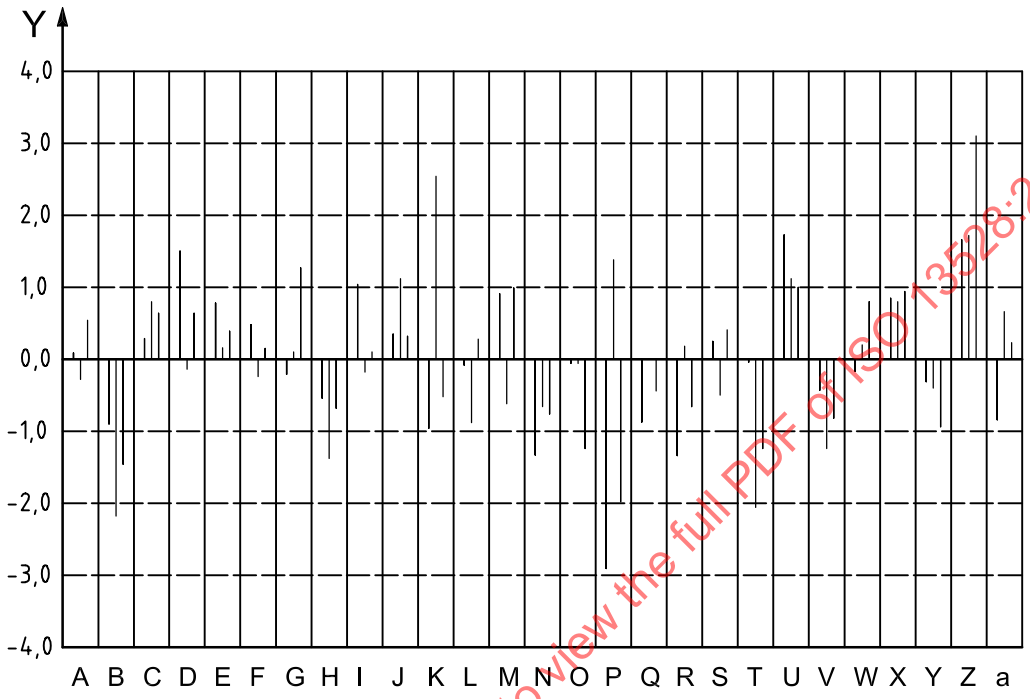
8.3.1 General

To prepare a bar-plot, collect the z-scores into a bar-plot as shown in Figure 9, in which the z-scores for each participant are grouped together. (The z-score is the same statistic as that referred to as the h-statistic in ISO 5725-2, and the bar-plots described here are the same graphs as the plots of h-statistics described in ISO 5725-2.)

Bar-plots are a suitable method of presenting the z-scores for a number of similar characteristics in one graph. They will reveal if there is any common feature in the z-scores for a participant, for example if he achieves several high z-scores indicating generally poor performance.

8.3.2 Example: Antibody concentrations

The z-scores from Table 7 are shown plotted as a bar-chart in Figure 9. From this graph, laboratories B and Z can see that they should look for a cause of bias that affects all three levels by approximately the same amount, whereas laboratories K and P can see that in their case the sign of the z-score depends on the type of antibody.



Key

Y z-score

NOTE "A" to "a" indicates the laboratory number.

Figure 9 — Bar-chart of z-scores (4,0 to –4,0) **for one round of a proficiency test in which the participants determined the concentrations of three allergen specific IgE antibodies**
(data from Table 7)

8.4 Bar-plots of standardized repeatability measurements

When replicate determinations are made in a round of a proficiency test, the results may be used to calculate a graph of k-statistics, as described in ISO 5725-2.

8.5 Youden Plot

8.5.1 General

When samples of two similar materials have been tested in a round of a proficiency scheme, the Youden Plot provides a very informative graphical method of studying the results. It is constructed by plotting the z-scores obtained on one of the materials against the z-scores obtained on the other material. A confidence ellipse, calculated as described in 8.5.2, is used as an aid to interpretation of the plot. A Youden Plot for the original data, laboratory biases, or percentage biases may be derived from the z-scores as explained in Note 1 below.

When a Youden Plot is constructed, interpret it as follows.

a) Inspect the plot for points that are well-separated from the rest of the data. If a laboratory is not following the test method correctly, so that its results are subject to bias, a point will be given far out along the major axis of the ellipse. Such a point can also occur if a laboratory suffers a large variation from time to time in the level of its results. Points far away from the major axis represent participants whose repeatability is poor.

b) Inspect the plot to see if there is evidence of a general relationship between the results for the two materials. If there is, then it shows that there is a cause of between-laboratory variation that is common to many of the laboratories, and provides evidence that the measurement method has not been adequately specified. Investigation of the test method may then allow the reproducibility of the method to be generally improved. The rank correlation test described below may be used to test if the relationship between the two materials is statistically significant. The rank correlation coefficient is preferred here to the correlation coefficient as the latter would be more sensitive to non-normality in the data.

8.5.2 **Confidence ellipse** (based on the method of Jackson [2])

8.5.2.1 General

Call the two materials A and B, and denote the results obtained on A by:

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

and those obtained on B by:

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

where p is the number of laboratories.

Calculate the averages and standard deviations of the two sets of data:

$$\bar{x}_{A,.}\, \bar{x}_{B,.} \quad s_A\, s_B$$

and the correlation coefficient $\hat{\rho}$. Calculate the z-scores for the two materials as:

$$z_{A,i} = \left(x_{A,i} - \bar{x}_{A,.}\right) / s_A \quad \text{where } i = 1, 2, \ldots, p \tag{31}$$

$$z_{B,i} = \left(x_{B,i} - \bar{x}_{B,.}\right) / s_B \quad \text{where } i = 1, 2, \ldots, p \tag{32}$$

and calculate the combined scores for the two materials:

$$z_{A,B,i} = \sqrt{z_{A,i}^2 - 2\hat{\rho}\, z_{A,i} z_{B,i} + z_{B,i}^2} \tag{33}$$

Define standardized variables as:

$$z_A = \left(x_A - \bar{x}_{A,.}\right) / s_A \tag{34}$$

$$z_B = \left(x_B - \bar{x}_{B,.}\right) / s_B \tag{35}$$

In terms of the standardized variables, the confidence ellipse may be written in terms of Hotelling's T^2:

$$z_A^2 - 2\hat{\rho} z_A z_B + z_B^2 = \left(1 - \hat{\rho}^2\right) T^2 \tag{36}$$

where

$$T^2 = 2\left\{(p-1)/(p-2)\right\} F_{(1-\alpha)}(2, p-1) \tag{37}$$

Here, $F_{(1-\alpha)}(2, p-1)$ is the tabulated $(1-\alpha)$-fractile of the F-distribution with 2 and $(p-1)$ degrees of freedom. The ellipse may be drawn on a graph that has the z-scores z_A and z_B as the axes by plotting a series of points for $-T \leqslant z_A \leqslant T$ with:

$$z_B = \hat{\rho}\, z_A \pm \sqrt{\left(1 - \hat{\rho}^2\right)\left(T^2 - z_A^2\right)} \tag{38}$$

NOTE 1 To plot the confidence ellipse on a graph with axes that show the original units of measurement, transform the above series of points back to the original units using:

$$x_A = \bar{x}_{A,.} + s_A \times z_A$$

$$x_B = \bar{x}_{B,.} + s_B \times z_B$$

To plot the confidence ellipse on a graph with axes that show laboratory biases D_A and D_B, transform the above series of points using:

$$D_A = s_A \times z_A$$

$$D_B = s_B \times z_B$$

To plot the confidence ellipse on a graph with axes that show percentage differences $D_{A\%}$ and $D_{B\%}$, transform the above series of points using:

$$D_{A\%} = 100 s_A \times z_A / x_A$$

$$D_{B\%} = 100 s_B \times z_B / x_B$$

The combined z-scores may be used as an aid to interpreting the Youden Plot. The highest combined z-scores correspond to the highest significance levels 100α % in the calculation of the confidence ellipse, so the combined z-scores may be used to identify the most extreme points on the Youden Plot. On occasion, it may be necessary to exclude one or more outlying points and re-calculate the ellipse: the combined z-scores may then be used as an aid to identifying the points to exclude.

NOTE 2 There is a need for a robust method of calculating the ellipse, but the details of such a method have not yet been worked out. The cut-off value may be calculated by noting that $(z_{A,B,i})^2 / (1-\hat{\rho}^2)$ has approximately the chi-squared distribution with 2 degrees of freedom, but the correction factor may have to be derived by simulation.

8.5.2.2 Example: Antibody concentrations

Table 10 shows data obtained by testing two similar samples for antibody concentrations, and the calculations required to derive the confidence ellipse. With p = 29 laboratories, and using a significance level of 100α % = 5 %, $F_{(1-\alpha)}(2, p-1)$ = 3,34. Hence T = 2,632 and, in terms of the standardized variables, the 95 % confidence ellipse may be written:

$$z_A^2 - 1{,}412\, z_A z_B + z_B^2 = 3{,}48 \tag{39}$$

The ellipse is shown, together with the points representing the z-scores, in Figure 10, together with the ellipses for probability levels of 100α % = 1 % and 0,1 %. The combined scores are shown in Table 10.

Inspection of Figure 10 reveals two laboratories (numbers 5 and 23) in the top right-hand quadrant. They have combined z-scores of 1,641 and 2,099. Laboratory 26 has a high z-score on Material B and a combined z-score of 2,059. After laboratories 5, 23 and 26, the laboratory that gives the next highest combined z-score is number 8 (combined score of 1,501).

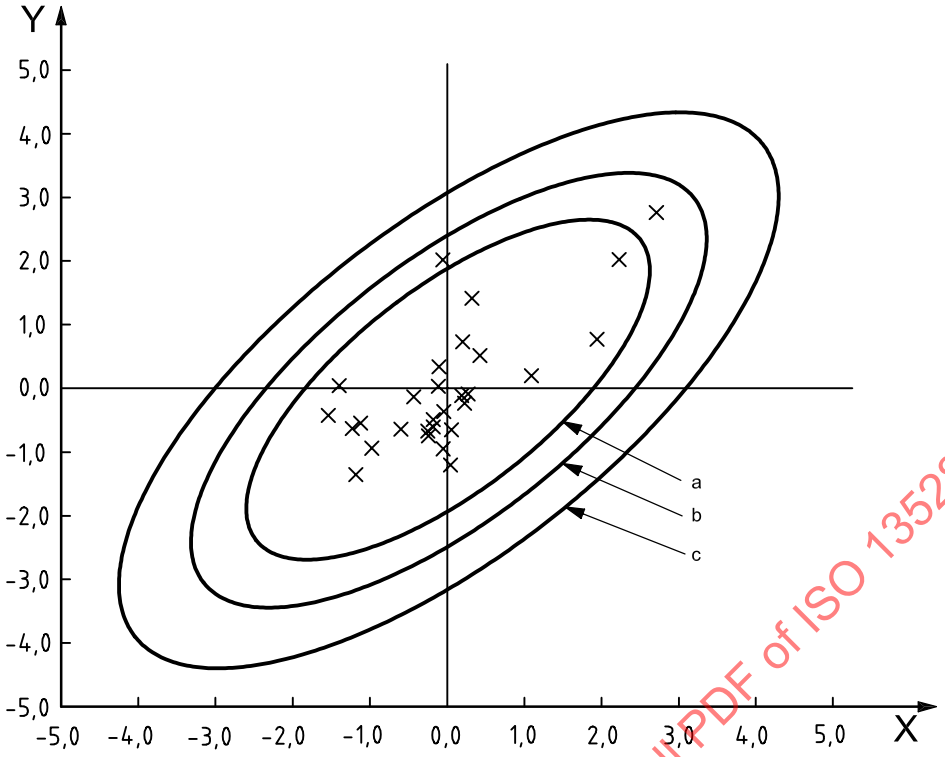
The points for laboratories 23 and 26 fall between the ellipses for the 5 % and 1 % probability levels, so it would be appropriate to treat their results as giving rise to “warning” signals, and to check where their results fall in the next round of the scheme.

Table 10 — Data and calculations on concentrations of antibodies for two similar allergens

Laboratory	Data		z-score		Combined score
	Allergen A	Allergen B	Allergen A	Allergen B	
i	$x_{A,i}$	$x_{B,i}$	$z_{A,i}$	$z_{B,i}$	$z_{A,B,i}$
1	12,95	9,15	0,427	0,515	0,370
2	6,47	6,42	−1,540	−0,428	1,275
3	11,40	6,60	−0,043	−0,366	0,336
4	8,32	4,93	−0,978	−0,942	0,737
5	18,88	13,52	2,228	2,023	1,641
6	15,14	8,22	1,092	0,194	0,965
7	10,12	7,26	−0,432	−0,138	0,349
8	17,94	9,89	1,942	0,770	1,501
9	11,68	4,17	0,042	−1,204	1,234
10	12,44	7,39	0,272	−0,093	0,344
11	6,93	7,78	−1,400	0,042	1,430
12	9,57	5,80	−0,599	−0,642	0,477
13	11,73	5,77	0,057	−0,652	0,693
14	12,29	6,97	0,227	−0,238	0,429
15	10,95	6,23	−0,180	−0,493	0,388
16	10,95	5,90	−0,180	−0,607	0,497
17	11,17	7,74	−0,113	0,028	0,134
18	11,20	8,63	−0,104	0,335	0,415
19	7,64	3,74	−1,185	−1,353	0,986
20	12,17	7,33	0,190	−0,114	0,282
21	10,71	5,70	−0,253	−0,676	0,529
22	7,84	6,07	−1,124	−0,549	0,833
23	20,47	15,66	2,710	2,762	2,099
24	12,60	11,76	0,321	1,415	1,210
25	11,37	4,91	−0,052	−0,949	0,913
26	11,36	13,51	−0,055	2,019	2,059
27	10,75	5,48	−0,241	−0,752	0,607
28	12,21	9,77	0,203	0,729	0,603
29	7,49	5,82	−1,230	−0,635	0,902
Average	11,54	7,66	0,00	0,00	
Standard deviation	3,29	2,90	1,00	1,00	
Correlation coefficient	0,706		0,706		

NOTE 1 The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

NOTE 2 The z-scores in this table have been calculated using non-rounded values of the averages and standard deviations, not using the rounded values shown at the bottom of the table.



Key

X z-score for allergen A

Y z-score for allergen B

a 5 % level.

b 1 % level.

c 0,1 % level.

Figure 10 — Youden Plot of z-scores from Table 10

8.5.3 Rank correlation test

8.5.3.1 General

The rank correlation test is carried out using Spearman's correlation coefficient as follows. Replace the observed results for Material A by their ranks (i.e. replace the lowest value by 1, the next lowest by 2, and so on, until the highest is replaced by p). Treat the observed results for Material B in the same way. Ties in the results are replaced by the average value of the ranks for the set of values that are tied. Calculate the correlation coefficient between the two series of ranks and compare the result with the tabulated value given in Table 11. If the calculated value is greater than the tabulated value, the relationship between the two variables is significant. The rank correlation coefficient may be calculated in the following simplified manner. Let

$$k_{A,1}, k_{A,2}, \ldots, k_{A,p}$$

represent the ranks of the laboratories for Material A and

$$k_{B,1}, k_{B,2}, \ldots, k_{B,p}$$

those for Material B. The rank correlation coefficient may be calculated as:

$$\rho_k = 1 - 6\sum\left(k_{A,i} - k_{B,i}\right)^2 \Big/ \left\{p\left(p^2 - 1\right)\right\} \qquad (40)$$

where the summation is over the p laboratories.

Table 11 — Critical values for the rank correlation coefficient

Number of data points	Critical values	
	5 % level	1 % level
8	0,738	0,881
9	0,683	0,833
10	0,648	0,794
11	0,623	0,818
12	0,591	0,780
13	0,566	0,745
14	0,545	0,716
15	0,525	0,689
16	0,507	0,666
17	0,490	0,645
18	0,476	0,625
19	0,462	0,608
20	0,450	0,591
21	0,438	0,576
22	0,428	0,562
23	0,418	0,549
24	0,409	0,537
25	0,400	0,526
26	0,392	0,515
27	0,385	0,505
28	0,377	0,496
29	0,370	0,487
30	0,364	0,478

8.5.3.2 Example: Antibody concentrations

An example of the calculation of the rank correlation coefficient is given in Table 12. The calculated rank correlation coefficient is 0,605 which exceeds the tabulated value in Table 11 of 0,487 for the 1 % level for 29 data points, so it may be concluded that the relationship apparent in Figure 10 is statistically significant.

Table 12 — Calculation of the rank correlation coefficient for the data from Table 10

Laboratory	Data		Rank		Difference
	Allergen A	Allergen B	Allergen A	Allergen B	
i	$z_{A,i}$	$z_{B,i}$	$k_{A,i}$	$k_{B,i}$	$\lvert k_{A,i} - k_{B,i} \rvert$
1	12,95	9,15	25	23	2
2	6,47	6,42	1	13	12
3	11,40	6,60	17	14	3
4	8,32	4,93	6	4	2
5	18,88	13,52	28	28	0
6	15,14	8,22	26	21	5
7	10,12	7,26	8	16	8
8	17,94	9,89	27	25	2
9	11,68	4,17	18	2	16
10	12,44	7,39	23	18	5
11	6,93	7,78	2	20	18
12	9,57	5,80	7	8	1
13	11,73	5,77	19	7	12
14	12,29	6,97	22	15	7
15	10,95	6,23	11,5	12	0,5
16	10,95	5,90	11,5	10	1,5
17	11,17	7,74	13	19	6
18	11,20	8,63	14	22	8
19	7,64	3,74	4	1	3
20	12,17	7,33	20	17	3
21	10,71	5,70	9	6	3
22	7,84	6,07	5	11	6
23	20,47	15,66	29	29	0
24	12,60	11,76	24	26	2
25	11,37	4,91	16	3	13
26	11,36	13,51	15	27	12
27	10,75	5,48	10	5	5
28	12,21	9,77	21	24	3
29	7,49	5,82	3	9	6
Sum of squared differences					1 605,50
$p(p^2-1)$					24 360
Calculated rank correlation coefficient					0,605
Tabulated critical value for the 1 % significance level					0,487

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

8.6 Plots of repeatability standard deviations

8.6.1 General

When n replicate measurements are made by the participants in a round of a proficiency testing scheme, the results may be used to produce a plot to identify any laboratories whose average and standard deviation are unusual. The graph is constructed by plotting the within-laboratory standard deviation s_i for each laboratory against the corresponding average x_i for the laboratory. Let

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

$\bar{S}$ = s^* the robust pooled value of $s_1, s_2, \ldots, s_p$, as calculated by Algorithm S

and assume that the data are Normally distributed. Under the null hypothesis that there is no difference between laboratories in the population values of either the laboratory means or the within-laboratory standard deviations, the statistic

$$\left(\sqrt{n}\,\frac{x_i - \bar{X}}{\bar{S}}\right)^2 + \left[\sqrt{2(n-1)}\,\ln\left(\frac{s_i}{\bar{S}}\right)\right]^2$$

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

$$s = \bar{S}\exp\left[\pm\frac{1}{\sqrt{2(n-1)}}\sqrt{\chi^2_{2;0,99} - \left(\sqrt{n}\,\frac{x-\bar{X}}{\bar{S}}\right)^2}\right] \quad (41)$$

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

$$x = \bar{X} - \bar{S}\sqrt{\frac{\chi^2_{2;0,99}}{n}} \quad \text{to} \quad \bar{X} + \bar{S}\sqrt{\frac{\chi^2_{2;0,99}}{n}} \quad (42)$$

NOTE The Circle Technique was introduced by van Nuland [6]. The method described used a simple Normal approximation for the distribution of the standard deviation that could give a critical region containing negative standard deviations. The method given here uses an approximation for the distribution of the standard deviation that avoids this problem, but the critical region is no longer a circle as in the original. Further, robust values are used for the central point in place of simple averages as in the original method.

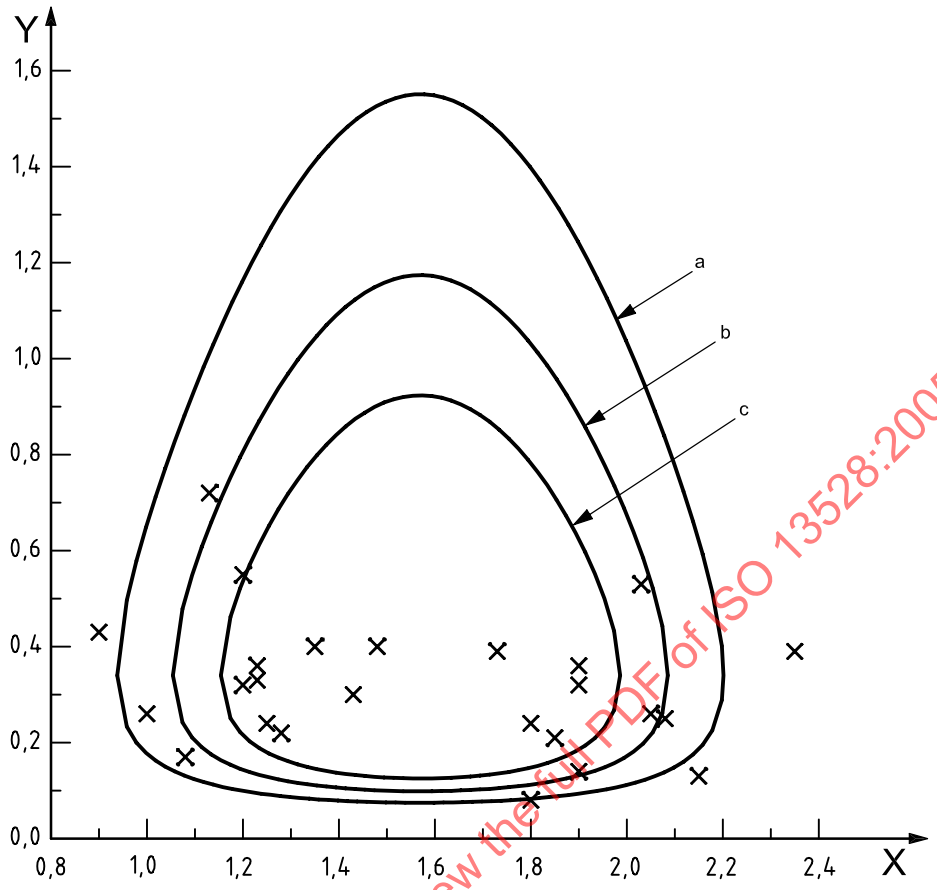
8.6.2 Example: Antibody concentrations

Table 13 shows the results of determining concentrations of a certain antibody in serum samples. Each laboratory made four replicate determinations, under repeatability conditions. The formulae given above are used to obtain the plot shown as Figure 11. The plot shows that several of the laboratories receive action or warning signals.

Table 13 — Concentrations of certain antibodies in serum samples
(four replicate determinations on one sample in each laboratory)

Laboratory	Average kU/l	Standard deviation kU/l
1	2,15	0,13
2	1,85	0,21
3	1,80	0,08
4	1,80	0,24
5	1,90	0,36
6	1,90	0,32
7	1,90	0,14
8	2,05	0,26
9	2,35	0,39
10	2,03	0,53
11	2,08	0,25
12	1,25	0,24
13	1,13	0,72
14	1,00	0,26
15	1,08	0,17
16	1,20	0,32
17	1,35	0,4
18	1,23	0,36
19	1,23	0,33
20	0,90	0,43
21	1,48	0,40
22	1,20	0,55
23	1,73	0,39
24	1,43	0,30
25	1,28	0,22
Robust average	1,57	
Robust standard deviation		0,34

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.



Key

X average

Y standard deviation

NOTE The data are numbers of units (U) in thousands (k) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

[a] 0,1 % level.

[b] 1 % level.

[c] 5 % level.

Figure 11 — Plot of standard deviations against averages for 25 laboratories
(data from Table 13)

8.7 **Split samples** (see ISO/IEC Guide 43-1:1997, A.3.1.2)

8.7.1 General

Split samples are used when it is necessary to carry out a detailed comparison of two laboratories. For example, if one laboratory is operated by a supplier, and the other by a customer, and the two organizations wish to ensure that the two laboratories are in agreement. Samples of several materials are obtained, representing a wide range of the property of interest, each sample is split into two parts, and each laboratory obtains some number (at least two) of replicate determinations on part of each sample.

On occasion, more than two laboratories may be involved, in which case one should be treated as a reference laboratory, and the others should be compared with it using the techniques described here.

The data from a split-sample experiment shall be used to produce graphs that display the variation between replicate measurements for the two laboratories and the differences between their average results for each sample. Further analysis will be dependent on deductions made from these graphs.

8.7.2 Example: Antibody concentrations

The concentration of certain antibodies in 21 serum samples were measured with radioimmunoassay methods in two laboratories denoted X and Y. In each laboratory all measurements were performed in duplicate in the same run. The concentrations obtained (in U/l) are presented in Table 14. As the measurement range is large, relative differences are relevant, so the data are transformed by taking logarithms to base e before the calculations are performed. The transformed data are shown in Table 15, and graphs showing the statistics from Table 15 are shown in Figures 12, 13 and 14.

From the graphs of ranges of replicate determinations, it appears that the variation between replicates for laboratory X is higher than for laboratory Y. Pooled values of these statistics are shown in Table 15, and could be compared using an F-test if it was of interest. Looking at the third graph, it can be seen that there is no obvious pattern or trend in the points. However, whereas the ranges of replicate determinations in Figures 12 and 13 are nearly all less than 0,2, many of the differences between laboratories in Figure 14 are much larger than this. This aspect requires investigation because it implies that the difference between the laboratories depends on the sample. The average difference between the laboratories may be calculated and is shown in Table 15. It may be used to give an indication of the importance of the difference between the laboratories, but it may not be used to predict the difference between the laboratories that might be obtained when analysing some subsequent sample. Thus with the transformed data, on average ln(Y) – ln(X) = 0,443, so Y/X = 1,6, indicating that laboratory Y obtains results, on average, higher than laboratory X by a factor of 1,6. However, with some samples the difference is much larger, and on others laboratory X obtains the higher results.

Table 14 — Concentrations of certain antibodies in 21 serum samples

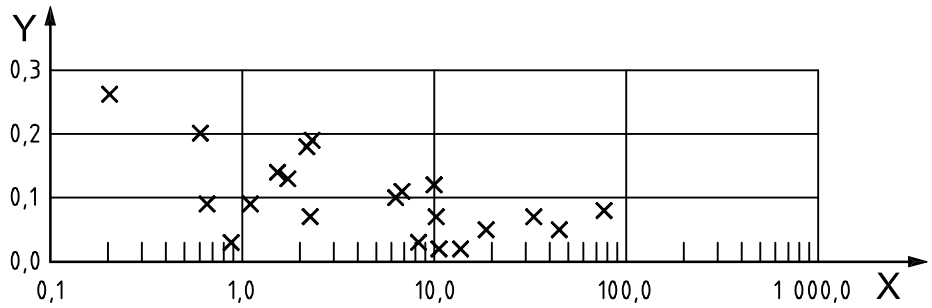
Sample	Laboratory X		Laboratory Y		Laboratory X	Laboratory Y	Laboratories X and Y
i	Replicate 1 U/l	Replicate 2 U/l	Replicate 1 U/l	Replicate 2 U/l	Average U/l	Average U/l	Average U/l
1	19,106	18,174	11,473	11,705	18,640	11,589	15,115
2	6,424	7,171	5,812	5,812	6,798	5,812	6,305
3	6,619	5,989	11,705	11,473	6,304	11,589	8,947
4	0,543	0,664	0,861	0,905	0,604	0,883	0,743
5	43,816	46,063	49,899	55,147	44,940	52,523	48,731
6	2,096	2,535	24,047	26,843	2,316	25,445	13,880
7	10,591	9,875	9,116	8,671	10,233	8,894	9,563
8	13,874	13,599	12,554	12,807	13,737	12,681	13,209
9	1,974	2,363	1,094	1,020	2,169	1,057	1,613
10	9,393	10,591	13,736	14,585	9,992	14,161	12,076
11	1,840	1,616	2,484	2,460	1,728	2,472	2,100
12	31,817	34,124	48,424	55,147	32,971	51,786	42,378
13	1,150	1,051	2,014	2,270	1,101	2,142	1,621
14	0,625	0,684	1,051	1,174	0,655	1,113	0,884
15	73,700	79,838	119,104	127,740	76,769	123,422	100,096
16	2,181	2,340	2,560	3,065	2,261	2,813	2,537
17	8,415	8,166	5,755	5,585	8,291	5,670	6,980
18	1,419	1,632	8,846	8,846	1,526	8,846	5,186
19	0,861	0,887	2,612	3,065	0,874	2,839	1,856
20	10,697	10,486	15,029	14,880	10,592	14,955	12,773
21	0,230	0,177	0,795	0,795	0,204	0,795	0,499

NOTE The data are numbers of units (U) per litre (l) of sample, where a unit is defined by the concentration of an international reference material.

Table 15 — ln(concentrations) and statistics for the data in Table 14

Sample	Laboratory X		Laboratory Y		Laboratory X	Laboratory Y	Difference Y – X
	Replicate 1	Replicate 2	Replicate 1	Replicate 2	Range	Range	
i	ln U/l	ln U/l	ln U/l	ln U/l	ln U/l	ln U/l	ln U/l
1	2,95	2,90	2,44	2,46	0,05	0,02	–0,475
2	1,86	1,97	1,76	1,76	0,11	0,00	–0,155
3	1,89	1,79	2,46	2,44	0,10	0,02	0,610
4	–0,61	–0,41	–0,15	–0,10	0,20	0,05	0,385
5	3,78	3,83	3,91	4,01	0,05	0,10	0,155
6	0,74	0,93	3,18	3,29	0,19	0,11	2,400
7	2,36	2,29	2,21	2,16	0,07	0,05	–0,140
8	2,63	2,61	2,53	2,55	0,02	0,02	–0,080
9	0,68	0,86	0,09	0,02	0,18	0,07	–0,715
10	2,24	2,36	2,62	2,68	0,12	0,06	0,350
11	0,61	0,48	0,91	0,90	0,13	0,01	0,360
12	3,46	3,53	3,88	4,01	0,07	0,13	0,450
13	0,14	0,05	0,70	0,82	0,09	0,12	0,665
14	–0,47	–0,38	0,05	0,16	0,09	0,11	0,530
15	4,30	4,38	4,78	4,85	0,08	0,07	0,475
16	0,78	0,85	0,94	1,12	0,07	0,18	0,215
17	2,13	2,10	1,75	1,72	0,03	0,03	–0,380
18	0,35	0,49	2,18	2,18	0,14	0,00	1,760
19	–0,15	–0,12	0,96	1,12	0,03	0,16	1,175
20	2,37	2,35	2,71	2,70	0,02	0,01	0,345
21	–1,47	–1,73	–0,23	–0,23	0,26	0,00	1,371
Pooled range					0,119	0,083	
Average difference between the laboratories							0,443

NOTE The data are numbers of units (U) per litre (l) of sample, where a unit is defined by the concentration of an international reference material. The pooled range is calculated according to Algorithm S in Annex C.

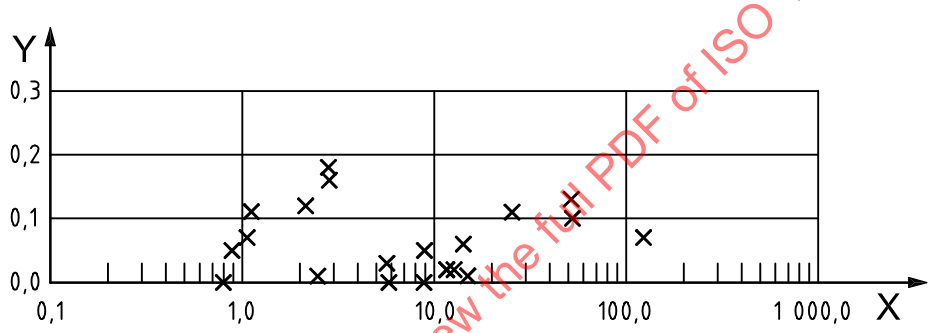


Key

X average concentration for laboratory X, %

Y range of ln(concentrations) for laboratory X

Figure 12 — Ranges of replicate determinations for laboratory X
(with the ranges calculated from the ln concentrations)

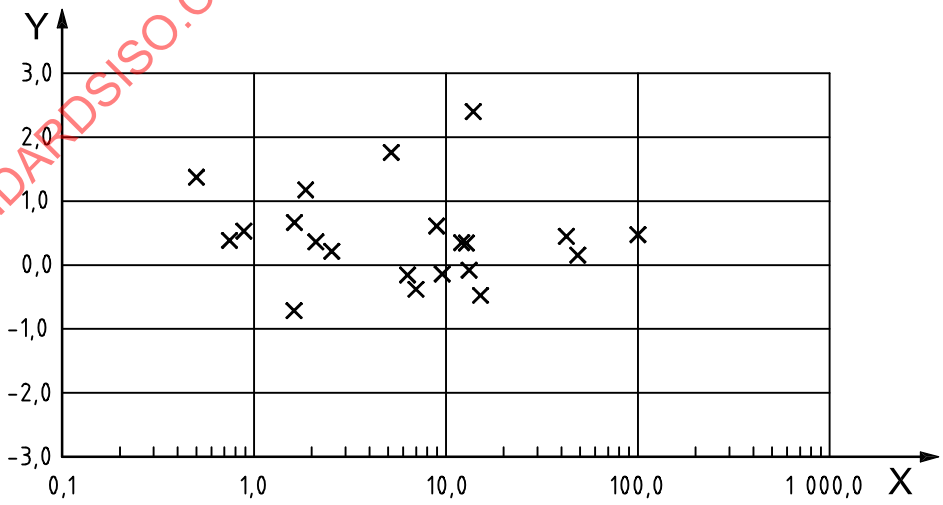


Key

X average concentration for laboratory Y, %

Y range of ln(concentrations) for laboratory Y

Figure 13 — Ranges of replicate determinations for laboratory Y
(with the ranges calculated from the ln concentrations)



Key

X average concentration for laboratories X and Y, %

Y difference in ln(concentrations) laboratory Y – laboratory X

Figure 14 — Differences between laboratory averages Y – X
(with the differences calculated from the ln concentrations)

9 Graphical methods for combining performance scores over several rounds of a proficiency testing scheme (see ISO/IEC Guide 43-1:1997, A.3.2)

9.1 Applications

When standardized scores are to be combined over several rounds, the coordinator shall consider preparing graphs, as described in 9.2 or 9.3. The use of these graphs, in which the scores for several rounds are combined, may allow trends, and other features of the results, to be identified that are not apparent when scores for each round are examined separately.

NOTE The use of "running scores", in which the scores obtained by a laboratory are combined over several rounds but not displayed graphically, is not recommended. The laboratory may have a fault that shows up with the test material used in one round but not in the others. A running score may hide this fault. The use of a running scores in the form of counts of the numbers of action and warning signals is described in 7.9.

9.2 Shewhart control chart for z-scores

9.2.1 General

To prepare this chart, the z-scores for a laboratory are plotted as individual points, with action and warning limits set at $\pm$ 2,0 and $\pm$ 3,0 in the style illustrated by Table 16 and Figure 15. When several characteristics are measured in each round, the z-scores for different characteristics may be plotted on the same graph, but the points for the different characteristics should be plotted using different plotting symbols and/or different colours. See ISO 8258 [4] for advice on plotting Shewhart charts.

The Shewhart control chart is an effective method of identifying problems that cause large erratic values of z-scores.

The rules for interpreting the Shewhart control chart are that an out-of-control signal is given when

a) a single point falls outside the action limits ($\pm$ 3,0),

b) two out of three successive points outside the same warning limit ($\pm$ 2,0).

When such a Shewhart control chart gives an out-of-control signal, the actions set out in 4.1 shall be initiated. Note that the standard deviation for proficiency assessment $\hat{\sigma}$ is not necessarily the standard deviation of the laboratory biases $x - X$, so the probability levels that are usually associated with the action and warning limits of a Shewhart control chart may not apply.

9.2.2 Example: Allergen concentrations

The data are shown in Table 16, and are plotted in Figure 15.