International Standard 5725

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

Precision of test methods — Determination of repeatability and reproducibility by inter-laboratory tests

Fidélité des méthodes d'essai — Détermination de la répétabilité et de la reproductibilité par essais interlaboratoires

First edition — 1981-04-01

UDC 519.2 : 620.1

Ref. No. ISO 5725-1981 (E)

Descriptors : tests, reproducibility, statistical analysis.

Price based on 42 pages

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards institutes (ISO member bodies). The work of developing International Standards is carried out through ISO technical committees. Every member body interested in a subject for which a technical committee has been set up 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.

Draft International Standards adopted by the technical committees are circulated to the member bodies for approval before their acceptance as International Standards by the ISO Council.

International Standard ISO 5725 was developed by Technical Committee ISO/TC 69, *Applications of statistical methods*, and was circulated to the member bodies in October 1977.

It has been approved by the member bodies of the following countries :

Australia	Hungary	Romania
Austria	India	South Africa, Rep. of
Belgium	Israel	Spain
Brazil	Italy	Sweden
Bulgaria	Japan	Switzerland
Canada	Kenya	Turkey
Czechoslovakia	Mexico	United Kingdom
Denmark	Netherlands	USSR
France	New Zealand	Yugoslavia
Germany, F. R.	Poland	

The member body of the following country expressed disapproval of the document on technical grounds :

USA

Printed in Switzerland

Contents

Precision of test methods — Determination of repeatability and reproducibility by inter-laboratory tests

0 Introduction

0.1 Tests performed on presumably "identical materials" (see 4.2) in presumably identical circumstances do not, in general, yield identical results. This is attributed to unavoidable random errors inherent in every test procedure; the factors that may influence the outcome of a test cannot all be completely controlled. In the practical interpretation of test data, this variability has to be taken into account. For instance, the difference between a test result and a value specified by contract may be within the scope of unavoidable random errors, in which case a true deviation from specification has not been established. Similarly, comparing test results from two batches of material will not indicate a fundamental quality difference if the difference between them can be attributed to inherent variation in the test procedure.

0.2 Many different factors (apart from error due to a lack of homogeneity of samples) may contribute to the variability of a test procedure, for example

a) the operator;

b) the instruments and equipment used;

c) the calibration of the equipment;

d) the environment (temperature, humidity, air pollution), etc.

The variability will be larger when the tests to be compared have been performed by different operators and/or with different instruments than when they have been carried out by a single operator using the same instruments. Hence, many different measures of variability are conceivable according to the circumstances under which the tests have been performed.

0.3 However, two extreme measures of variability, termed repeatability and reproducibility, have been found sufficient to deal with most practical cases. Repeatability refers to tests performed at short intervals (see 4.3) in one laboratory by one operator, using the same equipment, while reproducibility refers to tests performed in different laboratories, which implies different operators and different equipment. Under repeatability conditions, factors a) to d) listed in 0.2 are considered as constants and do not contribute to the variability, while under reproducibility conditions they vary and contribute to the variability of the test results.

1 Scope

This International Standard provides practical numerical definitions for the repeatability r and the reproducibility R of the results of a standard test method.

It discusses the implications of these definitions, and presents some practical rules for the interpretation of r and R.

It also describes the organization and analysis of inter-laboratory experiments for the numerical determination of r and R.

2 Field of application

This International Standard is exclusively concerned with test methods the results of which are expressed quantitatively.

This International Standard is primarily intended to be applied to test methods that have previously been standardized and that are used in different laboratories.

With slight modifications this International Standard may also be applied to test methods in use within a single laboratory (see 3.1.5) but this case has not been dealt with in this document.

Only the simplest type of experiment needed for estimating r and R is considered. This consists of tests made on samples of identical material sent to a number of different laboratories for testing.

This International Standard does not provide measures of the errors in estimating r and R (see 3.5).

Section one : General principles

3 Quantitative definitions of repeatability and reproducibility of a standard test method

3.1 For practical purposes, quantitative definitions are needed; those given below are according to ISO 3534.[1]

The **repeatability** r is the value below which the absolute difference between two single test results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory, and a short interval of time), may be expected to lie with a specified probability; in the absence of other indications, the probability is 95 %.

The **reproducibility** R is the value below which the absolute difference between two single test results obtained with the same method on identical test material, under different conditions (*different operators, different apparatus, different* laboratories and/or different time), may be expected to lie with a specified probability; in the absence of other indications, the probability is 95 %.

In the above and elsewhere in this International Standard, a single test result is the value obtained by applying the standard test method fully once to a single specimen, and as such may be the mean of two or more observations or the result of a calculation from a set of observations as specified by the method.

3.1.1 The definitions apply to continuous variables. When the test result is discrete or is rounded off, r and R are each the minimum value equal to or below which the absolute difference between two single test results is expected to lie with a specified probability (95 % in the absence of other indications).

3.1.2 The symbols r and R for the repeatability and reproducibility are already in general use. In ISO 3534, r is recommended for the correlation coefficient and R (or w) for the range of a single series of observations. There should, however, be no confusion as, when quoted in a standard method, the full wording repeatability r and reproducibility R should be used.

3.1.3 The statistical analysis of section three aims at the determination of r and R corresponding to a 95 % probability. Values for other probabilities can easily be derived from these, as explained in section four. If required, the probability level adopted can be attached as a subscript, for example r_{95}, R_{95}, or r_{99}, R_{99}.

3.1.4 The definition of repeatability in 3.1 applies to any test method within any laboratory. When a test method has been standardized, it may be expected that the repeatability will be, at least approximately, the same for all laboratories using the standard procedure; and the main purpose of this International Standard is to establish a standard experimental method for determining the repeatability of a standard test method. With slight modifications, however, the same type of experiment can also be used to determine the repeatability of a test method in use within a single laboratory. If so, it should always be stated clearly that the value of the repeatability obtained is only valid within the laboratory in question.

3.1.5 The terms reproducibility and repeatability as defined in 3.1 cover the conditions of maximum and minimum variation respectively. Other intermediate measures could be envisaged, for example the variability of results within a laboratory over a long period of time when recalibration may have occurred. Such measures have not been dealt with in this document.

3.2 The terms repeatability and reproducibility are used because they have been in common use for several years, but according to statistical terminology, r and R are critical differences at the 95 % probability level valid for two single test results obtained under repeatability or reproducibility conditions. Also, it is sometimes the practice to carry out two or more tests and a critical difference corresponding to the average of such tests may be preferred instead of the repeatability r or the reproducibility R as defined in 3.1. Critical differences valid under such modified conditions can all be derived from the values of r and R as defined in 3.1. The requisite formulae and conversion factors are given in section four.

3.3 r and R may be applied in a variety of ways. They can serve :

— to verify that the experimental technique of a laboratory is up to standard;

— to compare tests performed on a sample from a batch of material with a specification;

— to compare test results obtained by a supplier and a consumer on the same batch of material, etc.

Some of these various uses of repeatability and reproducibility are also discussed in section four.

3.4 Precision is a general term for the closeness of agreement between replicate test results. Thus, the repeatability r and the reproducibility R describe the precision of a given test method under two different circumstances of replication. A series of inter-laboratory trials organized with the specific purpose of determining r and R will therefore be referred to in this International Standard as a precision experiment.

3.5 As a consequence of the unavoidable random errors in the test results, the values of r and R derived from a precision experiment are estimated values. The method recommended in this International Standard has, however, been found to yield values sufficiently precise to satisfy practical requirements, provided that the laboratories employing the method for normal purposes are similar to those that participated in the precision experiment. The precision estimates should be re-estimated if at some future date evidence is available that the laboratories which participated in the original precision experiment were not representative of those currently using the test method.

4 Practical implications of the definitions

4.1 Standard test method

4.1.1 The definition of reproducibility in 3.1 refers to a standard test method and, as stated in clause 2, it is for these methods that this International Standard is primarily intended. This means that there must be a standard : that is a written document that lays down in full detail how the test should be carried out, including how the test specimen should be obtained and prepared. That standard must be applied in all the tests forming part of a precision experiment. The values of r and R derived from such an experiment should always be quoted as valid only for tests carried out according to that standard.

4.1.2 The existence of a standard implies the existence of a standardizing authority (such as ISO), within which there is a standards panel or working group responsible for the establishment of the standard under consideration.

4.1.3 In this International Standard, an essential distinction is made between standardization experiments carried out by the standards panel in order to establish the standard, and a precision experiment organized in order to determine the repeatability and reproducibility once the standard has been established.

The standardization experiments may provide information on the value of the repeatability and reproducibility but this information will not be used in the final determination of precision. It is assumed that r and R have to be estimated exclusively from the data resulting from a precision experiment specially organized for this purpose.

It is further assumed that the planning and organization of a precision experiment is a separate task to be entrusted to a precision panel. There is no reason why this should not be the same as the standards panel.

4.1.4 A precision experiment usually requires the cooperation of a larger number of laboratories and the collection of a larger number of test results than is needed in a standardization experiment. Hence, the standard test method is tried out on a larger scale than before and a precision experiment must also be considered as a final test concerning the adequacy of the standard. In particular, pronounced differences between the results reported by different laboratories may indicate that the standard is not yet sufficiently detailed and can possibly be improved. If so, this should be reported to the standards panel with a request for further investigation. [See 9.6, 17.2 b) and c), and 17.3.]

4.2 Identical material

4.2.1 According to the definitions of 3.1, tests to determine the repeatability and the reproducibility must be made on identical material. In most cases, the material on which a test is performed is either destroyed or undergoes a change. In reality, identical material therefore means that the tests are performed on samples taken from a homogeneous batch of material. The degree of homogeneity of the batch from which these samples are taken is then of great importance.

4.2.2 A fluid or a fine powder can be satisfactorily homogenized by stirring. If the material to be tested consists of a mixture of powders of different relative density or of different grain size, some care is needed because segregation may result from shaking, for instance during transport.

4.2.3 When the tests have to be performed on specimens of solid materials which cannot be homogenized — such as metals, rubber or textile fabrics — and when the tests cannot be repeated on the same test piece, then the variability among test pieces due to the heterogeneity of the material will be inseparable from the error variability of the test equipment, and will form an inherent part of both the repeatability and the reproducibility.

4.2.4 In practice, r and R are often used in order to compare batches of commercial material with a specification, or to make a comparison between two batches of material. It is then essential that any heterogeneity in such batches of commercial material be incorporated in the values of r and R. Whether or not this is the case will depend on the way the samples used in the precision experiment are prepared. The point should be carefully considered in planning these experiments.

4.2.5 When the tests have to be performed on discrete objects which are not altered by testing, the tests could, in principle at least, be carried out using the same set of objects in different laboratories. This, however, would necessitate circulating the same set of objects around many laboratories often situated far apart, in different countries or continents, with a considerable risk of loss or damage during transport.

4.2.6 In some circumstances, many of the details of this International Standard may need to be modified, but in a large proportion of cases, it should be possible for the essentials of this International Standard to be complied with.

4.2.7 Special precautions should be taken where samples are unstable. (See 9.3.)

4.3 Short intervals of time

According to the definition of 3.1, tests for the determination of repeatability have to be made under constant operating conditions. This must be interpreted as meaning that, during the time the tests are made, such factors as listed in 0.2 can be kept constant. In practice, tests under repeatability conditions should be conducted in as short a time as possible in order to minimize changes in these factors, which, particularly in the case of 0.2 d), cannot always be guaranteed constant. (See 10.4.1.)

5 Statistical model

5.1 Definition

For estimating the precision of a test method, it is useful to assume that every single test result y is the sum of three components :

$$y = m + B + e \qquad \ldots (1)$$

where, for the particular material tested, m is the average, B is a term representing the deviation of the laboratories from m and e is a random error occurring in every test.

Other models are sometimes used but it is considered that equation (1) will cover the majority of practical cases. (See 5.6.)

5.2 Average, m

5.2.1 The average m of the material tested will be called the level of the test property; different materials (for example different compositions of concrete) will correspond to different levels.

5.2.2 In some situations, the concept of a true value μ of the test property may hold good, for example the true concentration of a solution that is being titrated. The level m is, however, not necessarily equal to the true value; a difference $(m - \mu)$, when it exists, is called the bias of the test method.

When r or R is used to test the difference between two test results, a bias will have no influence and can be ignored. But when r or R are used to compare test results with a value specified in a contract or in a standard, a bias will have to be taken into account if the specification refers to the true value μ and not to the level m. If a true value exists and is known, the analysis of a precision experiment may indicate that there is a bias. (See 19.2.5.)

5.2.3 In many technical situations, however, the level of the test property is exclusively defined by the test method and the notion of an independent true value does not apply.

5.3 Term B in the model (5.1)

5.3.1 This term is considered to be constant during any series of tests performed under repeatability conditions, but to behave as a random variable in a series of tests performed under reproducibility conditions. The distribution of this variable is assumed to be approximately normal but, in practice, it is sufficient that it is unimodal. Its variance will be denoted by

$$\mathrm{var}(B) = \sigma_L^2$$

and called the between-laboratory variance.

σ_L^2 includes the between-operator and between-equipment variabilities.

5.3.2 In general, B can be considered as the sum

$$B = B_o + B_s$$

of a random component B_o and a systematic component B_s.

5.3.3 In a single laboratory, such factors as listed in 0.2 cannot be kept completely constant in the long run. Hence, within laboratories, long-term variabilities will exist larger than those accounted for by the repeatibility. These long term variations will contribute a random component B_o to B.

5.3.4 In addition, there may exist permanent systematic differences between laboratories. Serious systematic differences may result from misreading of the standard for the test method or from the use of inadequate equipment. They should be investigated and corrected, and are not considered as included in the term B. Unavoidably, however, some systematic differences will remain between different laboratories. These may be due to the use of different measuring instruments or working in different climatic conditions, but even without such gross differences, variation can arise from operator technique and also from one instrument to another of the same make due to manufacturing variations. These will all contribute a systematic component B_s to B.

5.3.5 If there are in all N laboratories likely to use the method at any time, B_s will take only N discrete values and the term B in the model (5.1) can only be considered as a random variable if either the systematic differences B_s are so small that they can be ignored, or else if the test results from which the reproducibility criterion is obtained were carried out by laboratories that can be considered as selected at random from all the laboratories likely to use the method.

5.3.6 Therefore, some caution is needed when the test results to be compared are always performed by the same two laboratories. The example on the determination of the softening point of pitch given in clause 22 provides an illustration of this in that the results from laboratory 11 are consistently lower, by about 4 °C, than those from laboratory 1.

5.4 Error term e in the model (5.1)

5.4.1 This term represents a random error occurring in every single test result. The distribution of this variable is assumed to be approximately normal but, in practice, it is sufficient that the distribution is unimodal. Within a single laboratory, its variance

$$\mathrm{var}(e) = \sigma_w^2$$

is called the within-laboratory variance.

5.4.2 It may be expected that σ_w^2 will vary between laboratories due to differences in the skills of the operators or in the quality of the equipment used. This International Standard assumes, however, that when a test method has been properly standardized, the differences between laboratories should be small so that it is justifiable to establish a common value for the within-laboratory variance valid for all laboratories using the standard method.

5.4.3 This common value, which is an average of the variances taken over the laboratories participating in the precision experiment, will be called the repeatability variance and be designated by

$$\overline{\text{var}}(e) = \sigma_r^2$$

5.5 Relation between the model (5.1), *r* and *R*

When the model (5.1) is adopted, the repeatability r and the reproducibility R are given by

$$r = f\sqrt{2}\,\sigma_r \quad \ldots (2)$$

$$R = f\sqrt{2}\sqrt{\sigma_L^2 + \sigma_r^2} = f\sqrt{2}\,\sigma_R \quad \ldots (3)$$

where $(\sigma_R^2 = \sigma_L^2 + \sigma_r^2)$ is called the reproducibility variance.

The coefficient $\sqrt{2}$ is derived from the fact that r and R refer to the difference between two single test results, and f is a factor whose value depends both on the number of test results available for estimating the variances σ_r^2 and σ_R^2, and on the shape of the distributions of the random components B and e in the model. However, if these distributions are approximately normal (in practice unimodal), the number of test results is not too small, and if the probability level is 95 %, the factor f will never differ much from the value 2 and the use of this value throughout is therefore recommended in this International Standard. (Taking into account variations in f would lead to considerable complications that would not effectively contribute to the practical value of r and R.)

Hence :

$$r = 2{,}83\,\sigma_r$$

$$R = 2{,}83\,\sigma_R$$

As the values of repeatability variance (σ_r^2) and the reproducibility variance (σ_R^2) are not known, their estimates s_r^2 and s_R^2 are used instead.

5.6 Suitability of the model (5.1)

It is clear that the model presented in 5.1 is an approximation that, by extensive experience, is known to satisfy practical requirements as a working hypothesis for designing the experiments and analysing the data. The point of view adopted in this International Standard is that the model is an acceptable approximation so long as the experimental requirements of section two are heeded and the statistical tests of section three do not yield significant results and thereby indicate its unsuitability. What action should be taken when these statistical tests indicate that the model is unsuitable will be discussed together with these tests.

6 Design of a precision experiment

6.1 One layout is as follows : samples from q batches of material, representing q different levels of the test property, are sent to p different laboratories, which are instructed to perform n tests under repeatability conditions at each level. These n tests are thus made on identical material and this type of experiment will be called a uniform-level experiment.

6.2 An alternative, preferred in certain cases (see 10.4.2), is the split-level experiment; each level is split into two sub-levels A and B, which are only slightly different. Each laboratory receives one sample from each of these sub-levels for testing.

6.3 These layouts are fully exemplified by the case studies in section five and will not be discussed here. Practical considerations in planning and execution are deferred to section two.

7 Analysis of the data

7.1 The analysis of the data produced by a precision experiment must be considered as a statistical problem to be entrusted to a statistical expert. (See 8.2 and 9.2.)

7.2 Three successive stages can be recognized, namely

a) a critical examination of the data in order to identify and treat outliers or other irregularities, and contingently to test the suitability of the model;

b) computation of preliminary values of r and R for each level separately;

c) establishment of final values of r and R including the establishment of a relation between r, R and m when the analysis indicates that they depend on the level m. If r and/or R are judged to be independent of m, the final values taken are the simple average over the levels.

7.2.1 As detailed in sub-clauses 14.8 to 14.11, the analysis of a precision experiment recommended in this International Standard first computes for each level estimates, s_r^2 and s_L^2, of the repeatability variance and the between-laboratory variance, as defined in 5.4.3 and 5.3.1 respectively, and from these derives the values of the repeatability r and the reproducibility R.

7.3 The analysis, especially stage a) described in 7.2, includes a systematic application of statistical tests. A great variety of statistical tests that might be used for the purpose of this International Standard is available from the literature.

In order to standardize the statistical analysis as far as possible, a judicious choice had to be made, and only a limited number of statistical tests, as explained in section three, has been incorporated in this International Standard.

Section two : Organization of an inter-laboratory precision experiment

General remark

The methods of operation within different organizations are not expected to be identical. Therefore, the contents of this section are only intended as a guide to be appropriately modified to cater for a particular situation.

8 Personnel requirements

8.1 Panel

The actual planning of the experiment should be the task of a panel of experts familiar with the test method and its application.

8.2 Statistical expert

At least one member of the panel should have experience in the statistical design and analysis of experiments.

8.3 Executive officer

The actual organization of the experiment should be entrusted to a single laboratory, and a member of the staff of that laboratory shall take full responsibility. He will be the executive officer.

8.4 Supervisors

A staff member in each of the participating laboratories should be made responsible for organizing the actual performance of the tests in keeping with instructions received from the executive officer, and for reporting the test results.

8.5 Operators

In each laboratory, the tests shall be carried out by one operator selected as representative of those likely to perform the tests in normal operations. He should be instructed by the supervisor as to the dates on which, and the order in which, the tests have to be carried out, but the instructions should not amplify the test method itself.

9 Tasks and problems

9.1 The following questions should be discussed by the panel :

a) Is a satisfactory standard for the test method available?

b) What is the range of levels encountered in practice?

c) How many levels should be used in the experiment? (See 10.1.)

d) What are suitable materials to represent these levels?

e) Should the material be specially homogenized before preparing the samples or should the heterogeneity in the material be included in the values of r and R? (See 10.3.)

f) What number n of replicates should be specified and what amount of material should be sent to the laboratories? (See 10.1.)

g) Should each laboratory be sent n separate samples for each level or one sample for n replicate tests? (See 10.3.) Or is a split-level experiment desirable? (See 10.4.2.)

h) Should the laboratories be sent additional material for practical exercises before the official tests are performed? (See 10.5.4.)

j) How many laboratories should be recruited to co-operate in the experiment? (See 10.1.)

k) How should the laboratories be recruited and what requirements should they satisfy? (See 10.2.)

m) Which are the details concerning the test method when the application is difficult? What kind of precisions are given to minimize these difficulties?

n) What instructions should be issued to the supervisors concerning the execution of the tests, and to how many significant figures should the test results be reported? (See 10.4.1 and 10.5.3.)

p) What information should be requested in addition to the numerical test results? (See 10.6.)

9.2 The task of the statistical expert is to contribute his specialized knowledge in designing the experiment, to analyse the data and to write a report for submission to the panel following the instructions contained in section three.

9.3 The task of the executive officer is to organize the experiment as planned by the panel, and in particular

a) to enlist the co-operation of the requisite number of laboratories and see to it that supervisors are appointed;

b) to organize and supervise the preparation of the materials and samples, and the despatch of the samples. For each level, a certain quantity of material should be set aside as a reserve stock;

c) to give special instructions when samples are unstable;

d) to draft instructions (including the interval of time between consecutive determinations) and circulate them to the supervisors early enough in advance for them to raise comments or queries;

e) to design suitable forms for the operator to use as a working record and for reporting the test results;

f) to collect the test results and prepare a table suitable for the statistical analysis.

9.4 The task of the supervisor is

a) to hand out the samples to the operators in keeping with the instructions of the executive officer;

b) to supervise the execution of the tests. The supervisor should not take part in performing the test;

c) to collect the test results, with any anomalies or difficulties experienced, and to report them to the executive officer.

9.5 The task of the operators is to perform the tests according to the standard test method and to report any anomalies and difficulties experienced (see 10.5.2 and 10.5.5).

9.6 The final task of the panel is to discuss the report by the statistical expert, establish final values for the repeatability and reproducibility, and decide if further actions are required for improving the standard for the test method or with regard to laboratories that have been rejected as outliers (see 11.6.4).

9.7 As 9.2 and 9.6 are considered to be the final stages of the statistical analysis, further details will be given in section three.

10 Comments on clauses 8 and 9

10.1 Number of laboratories and levels

No hard and fast rules can be laid down. The number of levels in a precision experiment should be chosen in relation to the range of levels to be covered, bearing in mind the cost of performing tests.

If the range of levels is very wide, r and R can be expected to depend on the level m and the use of at least 6 levels seems desirable in order to establish the relation between these quantities in a satisfactory manner.

For the example on the determination of the softening point of pitch given in clause 22 with a range of levels from 88 to 102 °C, the use of 4 levels may be considered as more than is strictly needed.

The number of laboratories should to some extent depend on the number of levels. It is recommended that the number of laboratories should never be less than 8, and if only a single level is of interest, the number of laboratories should preferably be higher, say 15 or more.

Regarding the value of n, the recommended figure is 2 except where it is customary to make a larger number of replicates, for example with certain simple physical tests.

10.2 Recruitment of participating laboratories

10.2.1 From a statistical point of view, the laboratories participating in a precision experiment should be chosen at random out of all laboratories likely to use the test method under investigation. Volunteers may not represent a realistic cross-section of laboratories.

However, in practice, other considerations may intervene; for example, the requirement that the participating laboratories should be evenly distributed over different continents or climatic regions.

The panel should decide the recruitment policy and the requirements for the participating laboratories.

10.2.2 In enlisting the co-operation of the requisite number of laboratories, their responsibility should be clearly stated. An example of a questionnaire that may be used for this purpose follows.

Questionnaire on inter-laboratory study

Title of method (copy attached) : . . . .
. . . .

1 Our laboratory wishes to participate in the co-operative testing of this method for precision data.

YES ☐ NO ☐

2 As a participant, we understand that

a) all essential apparatus, chemicals, and other requirements specified in the method must be available in our laboratory when the program begins;

b) specified "timing" requirements (such as starting date, order of testing specimens, and finishing date) of the program must be rigidly met;

c) the method must be strictly adhered to;

d) samples must be handled in accordance with instructions;

e) a qualified operator must perform the tests.

Having studied the method and having made a fair appraisal of our capabilities and facilities, we feel that we will be adequately prepared for co-operative testing of this method.

3 Comments : . . . .
. . . .

Signature : . . . .

Company or laboratory : . . . .

10.3 Heterogeneity of material

When the material to be tested is not homogeneous, it is important to prepare the samples in the manner prescribed by the method, preferably starting with one batch of commercial material for each level. Some modifications may be necessary to ensure that the amount of material available is sufficient to cover the experiment and keep a certain stock in reserve. For the samples at each level, n separate containers should be used where there is any danger of the material deteriorating when the container has once been opened, for example hygroscopic material, oxidation or loss of volatile components. In the case of unstable materials, special instructions on storage and treatment should be prescribed.

In general, it is recommended that the material used in a precision experiment and the range of materials to which r and R therefore apply be clearly specified.

10.4 Actual organization of the tests

10.4.1 With q levels and n replicates, each participating laboratory has to carry out qn tests. The performance of these tests should be organized and the operators instructed as follows :

a) All qn tests should be performed by one and the same operator using the same equipment throughout.

b) Each group of n tests belonging to one level must be carried out under repeatability conditions, that is, in a short interval of time and the same operator.

c) If, in the course of the tests, the operator should drop out through illness or some other unforeseen circumstances, another operator may complete the tests, but this must be reported with the test results.

d) It is not necessary that all qn tests be performed strictly within a short interval; the q groups of n tests may be carried out on different days.

e) It is essential that a group of n tests under repeatability conditions be performed independently as if they were n tests on different materials. As a rule, however, the operator will know that he is testing identical material. If it is feared that this knowledge may influence his test results, and consequently the repeatability variance, then a split-level experiment (see 10.4.2) shall be considered as the best procedure. Randomization of qn tests could be considered if it would not affect repeatability conditions.

10.4.2 An alternative method sometimes adopted when $n = 2$ is that of using split-levels. Instead of testing two samples which the operator has been told should be identical or of performing two tests on the same specimen of material, two series of p samples are prepared at slightly different levels m_A and m_B, where $m_A - m_B$ is a small quantity, and each of the p laboratories receives one sample of series A and one of series B for testing. Adoption of this method may be considered when it is feared that the operator, when using identical samples in carrying out his second test, may be influenced by the result of his first test.

The split-level experiment requires a slight modification in the statistical analysis which will be discussed in section three. Also, it should be clearly distinguished which test result belongs to series A and which to series B; they cannot be interchanged as can tests on identical material.

The values of r and R derived from a split-level experiment will be taken to be valid for the mean level $m = (m_A + m_B)/2$.

10.4.3 It may be necessary to limit the time that should be allowed to elapse between the day the samples are received and the day the tests are performed.

10.4.4 Any preliminary checking of equipment should be as laid down in the standard method.

10.4.5 All samples should be clearly labelled with the name of the experiment and a sample identification.

10.5 Instructions to operators

10.5.1 Before performing the tests, the operators should receive no instructions other than those contained in the standard test method; these should suffice.

10.5.2 The operators should, however, be asked to comment on this standard and in particular to state whether the instructions contained in it are sufficiently unambiguous and clear. Ambiguities may, for example, creep in when the standard has been translated into a number of different languages.

10.5.3 It is desirable that all participating laboratories report their test results to the same number of significant figures and the supervisors should be instructed accordingly. In commercial practice, the test results are often rounded rather crudely and in a precision experiment, it may be advisable to use one more significant figure than is customary or prescribed in the Standard.

When r depends on the level m, different rules for rounding may be needed for different levels.

10.5.4 An operator will not as a rule achieve normal precision when he caries out a test for the first time or after a long interval. In that case, the operators should be instructed to carry out a few unofficial tests in order to gain experience before they start testing the official samples of the precision experiment. Whether this is needed should be decided by the panel or by the supervisors; material for such preliminary tests should be supplied by the executive officer.

10.5.5 The operators should be told to report any occasions on which they are not able to follow their instructions or on which they accidentally failed to keep to the instructions. They should also be told that it is better to report a mistake than to adjust the results, because one or two missing results will not spoil the experiment and may indicate a deficiency in the standard.

10.6 Reporting the test results

The supervisor of each laboratory should write a full report on the tests; this report should contain the following particulars :

a) the final test results, taking particular care to avoid transcription and typing errors, for example by using photocopies of the operator's results;

b) if any, the original observations or readings from which the final results were derived;

c) comments by the operator on the standard for the test;

d) information about irregularities or disturbances that may have occurred during the test;

e) the date(s) on which the samples were received;

f) the date(s) and time(s) on which they were tested;

g) information about the equipment used, when this is considered relevant;

h) other relevant information.

Section three : Statistical analysis of results of an inter-laboratory experiment

11 Preliminary considerations

11.1 Statistical expert

The analysis of the test results produced by a precision experiment is the task of a statistical expert who is a member of the panel and has taken part in planning the experiment. (See 8.2 and 9.2.)

11.2 Cells

Each combination of a laboratory and a level will be called a cell of the precision experiment. In the ideal case, the results of an experiment with p laboratories and q levels will consist of a table with pq cells each containing n replicate results, that can all be used for computing the repeatability r and the reproducibility R. This ideal situation is not, however, always attained in practice. Departures occur due to redundant data, missing data, and outliers.

11.3 Redundant data

Sometimes a laboratory may carry out more than the n replicates officially prescribed. In that case, the supervisor (see 8.4 and 9.4) must report all results, why this was done and which are the correct test results. If the answer is that they are all equally valid, they can all be taken into account by using the computational procedure of 14.10.

11.4 Missing data

In other cases, some of the test results may be missing, for example due to the loss of a sample, a slip in performing the test, etc. The analysis recommended in clause 16 is such that completely empty cells can simply be ignored, while partly empty cells can be taken into account by the computational procedure of 14.10. The reasons for missing test results should be given in the supervisor's report.

If one of the two test results in a cell of a split-level experiment (see 10.4.2) is missing, the test result available must be discarded and the cell must be treated as an empty one.

11.5 Outliers

Outliers are entries among the original test results, or in the tables derived from them, that deviate so much from comparable entries in the same table that they are considered as irreconcilable with the other data. Experience has taught that outliers cannot always be avoided and have to be taken into consideration, but great care must be exercised in investigating them.

11.6 Recommended practice for investigating outliers

11.6.1 For this purpose, this International Standard recommends the use of Cochran's maximum variance test (see clause 12) and Dixon's outlier test (see clause 13) in combination with the following procedure :

$P \geqslant 5$ %, that is, Cochran's or Dixon's test statistic is less than its 5 % critical value : the item tested is accepted as correct; the test is said to be statistically insignificant.

5 % $> P \geqslant$ 1 %, that is, the test statistic lies between its 5 % and 1 % critical values : the item tested is called a straggler and is marked with a single asterisk; the test is said to be statistically significant.

$P <$ 1 %, that is, the test statistic is greater than its 1 % critical value : the item is called a statistical outlier and is marked with a double asterisk; the test is said to be statistically highly significant.

P is the probability of the observed value of the test statistic.

The 5 % and 1 % critical values for Cochran's and Dixon's tests are given in tables 1 and 2 (pages 17 and 18).

11.6.2 Sometimes the actual application of these statistical tests may be omitted or other statistical tests may be chosen because a statistical expert will see from a cursory examination of the data, for example from a graphical presentation, that the test will yield either a non-significant or a highly significant result. In case of any doubt, however, the test should always be applied.

11.6.3 It is next investigated whether the stragglers and/or statistical outliers can be explained by some technical error, for instance a slip in performing the test, a computational error, a clerical error in transcribing a test result, the analysis of a wrong sample, etc. When a reasonable explanation can be construed, and preferably confirmed by additional enquiries, the item is considered as a real outlier, that does not belong to the experiment proper, and is corrected or discarded in keeping with the explanation obtained.

11.6.4 When several unexplained stragglers and/or statistical outliers occur at different levels within the same laboratory, that laboratory may be considered as an outlier, having too high a within-laboratory variance, or too large a systematic error in the level of its test results, or both. It may then be reasonable to discard some or all the data from such an outlying laboratory. This International Standard does not provide a statistical test by which suspected laboratories can be judged. The primary decision should be the responsibility of the statistical expert, but all rejected laboratories must be reported to the panel for further action. An example of an outlying laboratory occurs in the case study of clause 23.

11.6.5 When any stragglers and/or statistical outliers remain that have not been explained or rejected as belonging to an outlying laboratory, the stragglers are retained as correct items, and the statistical outliers are discarded, unless the statistician for good reasons decides to retain them. In any case, the statistician must report to the panel.

11.7 Computation of r and R

The computation of the repeatability r and the reproducibility R is carried out, for each level separately, with the data remaining after elimination or correction of the stragglers and/or outliers (see clause 14).

11.8 Functional relation between r, R and m

Provided there are several levels and a functional relation between r (and/or R) and m is expected (see 15.1), it is then investigated whether r (and/or R) depends on m and if so, what is the relationship between these quantities.

11.9 Notation, definitions and basic formulae

11.9.1 Ideal case

As stated in 11.2, the ideal case is p laboratories L_i (i = 1, 2, ..., p), q levels M_j (j = 1, 2, ..., q) and n replicates per $L_i M_j$ combination, with a total npq results of the test. As a result of the existence of redundant (see 11.3), missing (see 11.4) or outlying (see 11.5) results, or outlying laboratories (see 11.6.4), this ideal situation is not always attained. Under these conditions, the notations given in 11.9.2 to 11.9.6 will be used in the remainder of this International Standard, refering to basic tables A, B, C hereafter.

11.9.2 Original test results (table A)

11.9.2.1 Case of a uniform-level experiment

n_{ij} is the number of results in cell $L_i M_j$,

y_{ijk} is any one of these results (k = 1, 2, ..., n_{ij}).

When all the results from one or more laboratories are eliminated at a level j (either because these laboratories did not conduct the tests at this level, or because they were considered to be "outlying laboratories" — see 11.6.4), the number of laboratories for this level is designated by p_j.

11.9.2.2 Case of a split-level experiment

y_{ijA} and y_{ijB} are the results obtained, respectively at sub-levels A and B, level j, laboratory i. The notation p_j is also applicable to this case, where appropriate.

Table A — Original test results

Uniform-level experiment

Laboratory \ Level	1	2		j		q
1						
2						
i				y_{ij1} ... y_{ijk}		
p						

Cell (i, j) contains n_{ij} results y_{ijk} (k = 1, 2, ... n_{ij}).

Split-level experiment

Laboratory \ Level	1 A	1 B	2 A	2 B		j A	j B		q A	q B
1										
2										
i						y_{ijA}	y_{ijB}			
p										

Table B — Measures of cell spread

Uniform-level experiment

Laboratory \ Level	1	2	j		q
1					
2					
i			s_{ij} or w_{ij}		
p					

s_{ij} = cell standard deviation, or if n = 2 for all the cells, use w_{ij} = cell range.

Split-level experiment

Laboratory \ Level	1	2		j		q
1						
2						
i				d_{ij}		
p						

$d_{ij} = y_{ij\mathrm{A}} - y_{ij\mathrm{B}}$

Table C — Cell averages

Laboratory \ Level	1	2		j		q
1						
2						
i				$\bar{y}_{ij}$		
p						

$\bar{y}_{ij}$ = cell average

11.9.3 Measures of cell spread (table B)

These are derived from table A (see 11.9.2) and table C (see 11.9.4) as follows :

11.9.3.1 Case of a uniform-level experiment

For the general case, use the intra-cell standard deviations s_{ij}, given by equation (4)

$$s_{ij} = \sqrt{\frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}$$

$$= \sqrt{\frac{1}{n_{ij}-1}\left[\sum_{k=1}^{n_{ij}} (y_{ijk}^2) - \frac{1}{n_{ij}}\left(\sum_{k=1}^{n_{ij}} y_{ijk}\right)^2\right]} \quad \ldots (4)$$

The standard deviation should be expressed with one more significant figure than the results in table A.

For the particular case where all $n_{ij} = n = 2$, use the cell range

$$w_{ij} = | y_{ij1} - y_{ij2} | = s\sqrt{2} \quad \ldots (5)$$

without regard for sign.

11.9.3.2 Case of a split-level experiment

The cell difference d_{ij} is given by equation (6)

$$d_{ij} = y_{ijA} - y_{ijB} \quad \ldots (6)$$

taking the sign into account.

11.9.4 Cell averages (for the two types of experiment) (table C)

These are derived from table A as follows :

$$\bar{y}_{ij} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk} \quad \ldots (7)$$

The cell averages should be given with one more significant figure than the test results in table A.

11.9.5 Repeatability variance s_r^2 and between-laboratory variance s_L^2

For a given level j, the values of s_r^2 and s_L^2 are given by the following equations where, for convenience, the index j has been dropped.

11.9.5.1 Case of a uniform-level experiment

$$s_r^2 = \frac{\sum_{i=1}^{p} (n_i - 1)\, s_i^2}{\left(\sum_{i=1}^{p} n_i\right) - p}$$

$$s_L^2 = \frac{\frac{1}{p-1}\sum_{i=1}^{p} n_i (\bar{y}_i - \bar{\bar{y}})^2 - s_r^2}{\bar{\bar{n}}}$$

$$\text{with } \bar{\bar{y}} = \frac{\sum_{i=1}^{p} n_i \bar{y}_i}{\sum_{i=1}^{p} n_i}$$

$$\text{and } \bar{\bar{n}} = \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i - \frac{\sum_{i=1}^{p} n_i^2}{\sum_{i=1}^{p} n_i}\right]$$

Formulae used for numerical computation are given in 14.10.2. For the case where all $n_i = n$, the previous formulae simplify to

$$s_r^2 = \frac{1}{p}\sum_{i=1}^{p} s_i^2$$

$$s_L^2 = \frac{1}{p-1}\sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2 - \frac{s_r^2}{n}$$

$$\text{with } \bar{\bar{y}} = \frac{1}{p}\sum_{i=1}^{p} \bar{y}_i$$

Formulae used for numerical computation are given in 14.9.2.

For the particular case where all $n_i = n = 2$, the cell range $w_i = \sqrt{2}\, s_i$ are used, giving

$$s_r^2 = \frac{1}{2p}\sum_{i=1}^{p} w_i^2$$

$$s_L^2 = \frac{1}{p-1}\sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2 - \frac{s_r^2}{2}$$

Formulae used for numerical computation are given in 14.8.2.

11.9.5.2 Case of a split-level experiment

$$s_r^2 = \frac{1}{2} \cdot \frac{1}{p-1} \sum_{i=1}^{p} (d_i - \bar{d})^2$$

with $\bar{d} = \frac{1}{p} \sum_{i=1}^{p} d_i$

$$s_L^2 = \frac{1}{p-1} \sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2 - \frac{s_r^2}{2}$$

with $\bar{\bar{y}} = \frac{1}{p} \sum_{i=1}^{p} \bar{y}_i$

Formulae used for numerical computation are given in 14.11.2.

11.9.6 Simplified notations used in clauses 12, 13 and 14

Clauses 12 and 13 concern statistical tests and clause 14 relates to procedures for calculating r and R, which are applied separately at each level (fixed j); in these clauses, for reasons of clarity of layout, the index j will be omitted in the notations defined above, when this index is not indispensable.

11.9.7 Corrected or rejected data

As on the basis of the tests outlines in 11.6, some of the data may be corrected or rejected, the values of y_{ijk}, n_{ij} and p_j used for the final determination of r and R may be different from the values referring to the original test results as recorded in tables A, B and C. Hence, in reporting the final values of r and R, it should always be stated what data, if any, have been corrected or discarded.

12 Cochran's maximum variance test

12.1 As explained in 5.4.2, this International Standard assumes that between laboratories only small differences exist in the within-laboratory variances. As experience shows that this condition is not always satisfied, a test has been included to investigate the validity of this assumption. Three tests could be used for this purpose, namely

a) Bartlett's variance homogeneity test;

b) Hartley's variance ratio test;

c) Cochran's maximum variance test.

All three are fully explained in the literature.[4]

The first two tests, however, cannot be applied when one of the variances in a set is zero, which may easily happen as a result of rounding and of the small number of test results on which the variances are based. Moreover, these tests, even if no zeros occur, are very sensitive against the value of the smallest variance which, again due to rounding, is unreliable. For these reasons, only Cochran's test has been retained.

Cochran's test applies only to uniform-level experiments as it is of homogeneity of variance; in a split-level experiment, Dixon's outlier test (see clause 13) must be applied to the cell differences d_i.

12.2 Given a set of p standard deviations s_i, all computed from the same number n of replicate test results, Cochran's criterion C is given by equation (8)

$$C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2} \qquad \ldots (8)$$

In the case of 2 replicates, the ranges w_i can be used instead of the standard deviations. Cochran's criterion then is given by equation (9)

$$C = \frac{w_{max}^2}{\sum_{i=1}^{p} w_i^2} \qquad \ldots (9)$$

In these expressions, s_{max} and w_{max} stand for the highest values in the set. If the test is significant, s_{max} (or w_{max}) is classified as straggler or statistical outlier according to the procedure of 11.6.1. Critical values for Cochran's criterion at the 5 % and 1 % levels are given for p = 2 to 40 and n = 2 to 6 in table 1.

12.3 Cochran's criterion must be applied to table B at each level separately.

12.4 As stated above, Cochran's criterion applies strictly only when all standard deviations are derived from the same number of test results obtained under conditions of repeatability. In actual cases, this number may vary due to redundant, missing or discarded data. This International Standard assumes, however, that in a properly organized experiment, such variations in the number of test results per cell will be limited and can be ignored, Cochran's criterion being applied using for n the number of results occurring in the majority of cells.

12.5 Cochran's criterion tests only the highest value in a set of standard deviations or ranges and is therefore a one-sided outlier test. Variance heterogeneity may, of course, also manifest itself in some of the standard deviations being comparatively too low. However, small values of standard deviation or range may be very strongly influenced by the degree of rounding of the original test results and are for that reason not very reliable. Besides, it does not seem recommendable to reject the data of some laboratory because it has accomplished a higher precision in its test results than the other laboratories. Hence, Cochran's criterion is considered adequate.

12.6 A critical examination of table B may sometimes reveal that the standard deviations for a particular laboratory are at all or at most levels lower than those for other laboratories. This may indicate that the laboratory works with a lower repeatability than the other laboratories, which in turn may be due either to an incorrect application of the standard test method or to a better technique and equipment. If this occurs, it should be reported to the panel, which should decide whether the point is worthy of more detailed investigation.

12.7 When there is more than one high value among a set of standard deviations suspect as possible outliers, Cochran's test may become insensitive. It may happen that the highest value produces an insignificant test statistic while the second or third highest turn out to be stragglers or statistical outliers when the highest or the two highest values are disregarded. In that case, the two or three highest standard deviations should all be marked as stragglers or statistical outliers as the case may be.

Thus, in view of the present lack of a statistical test designed for testing several outliers together, repeated application of Cochran's test disregarding the higher standard deviations in order of magnitude is proposed as a helpful tool. This test is not, however, designed for this purpose and great caution should be exercised in drawing conclusions.

In particular, when this technique reveals several stragglers and/or statistical outliers only within one of the levels, this may be purely accidental and not really significant. On the other hand, if several stragglers and/or statistical outliers are found at different levels within one laboratory, this may be a strong indication that that laboratory's within-laboratory variance is too high and that its experimental technique can and should be improved.

Table 1 — Critical values for Cochran's maximum variance test[2]

p	$n = 2$		$n = 3$		$n = 4$		$n = 5$		$n = 6$	
	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %
2	—	—	0,995	0,975	0,979	0,939	0,959	0,906	0,937	0,877
3	0,993	0,967	0,942	0,871	0,883	0,798	0,834	0,746	0,793	0,707
4	0,968	0,906	0,864	0,768	0,781	0,684	0,721	0,629	0,676	0,590
5	0,928	0,841	0,788	0,684	0,696	0,598	0,633	0,544	0,588	0,506
6	0,883	0,781	0,722	0,616	0,626	0,532	0,564	0,480	0,520	0,445
7	0,838	0,727	0,664	0,561	0,568	0,480	0,508	0,431	0,466	0,397
8	0,794	0,680	0,615	0,516	0,521	0,438	0,463	0,391	0,423	0,360
9	0,754	0,638	0,573	0,478	0,481	0,403	0,425	0,358	0,387	0,329
10	0,718	0,602	0,536	0,445	0,447	0,373	0,393	0,331	0,357	0,303
11	0,684	0,570	0,504	0,417	0,418	0,348	0,366	0,308	0,332	0,281
12	0,653	0,541	0,475	0,392	0,392	0,326	0,343	0,288	0,310	0,262
13	0,624	0,515	0,450	0,371	0,369	0,307	0,322	0,271	0,291	0,246
14	0,599	0,492	0,427	0,352	0,349	0,291	0,304	0,255	0,274	0,232
15	0,575	0,471	0,407	0,335	0,332	0,276	0,288	0,242	0,259	0,220
16	0,553	0,452	0,388	0,319	0,316	0,262	0,274	0,230	0,246	0,208
17	0,532	0,434	0,372	0,305	0,301	0,250	0,261	0,219	0,234	0,198
18	0,514	0,418	0,356	0,293	0,288	0,240	0,249	0,209	0,223	0,189
19	0,496	0,403	0,343	0,281	0,276	0,230	0,238	0,200	0,214	0,181
20	0,480	0,389	0,330	0,270	0,265	0,220	0,229	0,192	0,205	0,174
21	0,465	0,377	0,318	0,261	0,255	0,212	0,220	0,185	0,197	0,167
22	0,450	0,365	0,307	0,252	0,246	0,204	0,212	0,178	0,189	0,160
23	0,437	0,354	0,297	0,243	0,238	0,197	0,204	0,172	0,182	0,155
24	0,425	0,343	0,287	0,235	0,230	0,191	0,197	0,166	0,176	0,149
25	0,413	0,334	0,278	0,228	0,222	0,185	0,190	0,160	0,170	0,144
26	0,402	0,325	0,270	0,221	0,215	0,179	0,184	0,155	0,164	0,140
27	0,391	0,316	0,262	0,215	0,209	0,173	0,179	0,150	0,159	0,135
28	0,382	0,308	0,255	0,209	0,202	0,168	0,173	0,146	0,154	0,131
29	0,372	0,300	0,248	0,203	0,196	0,164	0,168	0,142	0,150	0,127
30	0,363	0,293	0,241	0,198	0,191	0,159	0,164	0,138	0,145	0,124
31	0,355	0,286	0,235	0,193	0,186	0,155	0,159	0,134	0,141	0,120
32	0,347	0,280	0,229	0,188	0,181	0,151	0,155	0,131	0,138	0,117
33	0,339	0,273	0,224	0,184	0,177	0,147	0,151	0,127	0,134	0,114
34	0,332	0,267	0,218	0,179	0,172	0,144	0,147	0,124	0,131	0,111
35	0,325	0,262	0,213	0,175	0,168	0,140	0,144	0,121	0,127	0,108
36	0,318	0,256	0,208	0,172	0,165	0,137	0,140	0,118	0,124	0,106
37	0,312	0,251	0,204	0,168	0,161	0,134	0,137	0,116	0,121	0,103
38	0,306	0,246	0,200	0,164	0,157	0,131	0,134	0,113	0,119	0,101
39	0,300	0,242	0,196	0,161	0,154	0,129	0,131	0,111	0,116	0,099
40	0,294	0,237	0,192	0,158	0,151	0,126	0,128	0,108	0,114	0,097

p = the number of laboratories at a given level.

n = number of results per cell (see 12.4).

These critical values are those given in table A17[2] rounded and supplemented by interpolation, using the fact that the critical values are closely approximated by a linear function in $1/\sqrt{p}$.

This table contains critical values for n = 7, 8, ..., as well.

13 Dixon's outlier test

13.1 Given a set of data $z(h)$, $h = 1, 2, 3, \ldots, H$, arranged in order of magnitude, then Dixon's test uses the following test statistics :

H	Test statistic
3 to 7	Q_{10} = the larger of $\dfrac{z(2) - z(1)}{z(H) - z(1)}$ and $\dfrac{z(H) - z(H-1)}{z(H) - z(1)}$
8 to 12	Q_{11} = the larger of $\dfrac{z(2) - z(1)}{z(H-1) - z(1)}$ and $\dfrac{z(H) - z(H-1)}{z(H) - z(2)}$
13 or more	Q_{22} = the larger of $\dfrac{z(3) - z(1)}{z(H-2) - z(1)}$ and $\dfrac{z(H) - z(H-2)}{z(H) - z(3)}$

Critical values of these test statistics at the 5 % and 1 % level and for H = 3 to 40 are reproduced in table 2.

13.2 In analysing a precision experiment, Dixon's test can be applied

a) to the test results within a cell of table A when $n_{ij} \geqslant 3$, but this procedure should only be used where Cochran's test has suggested an outlier or a straggler, in order to see whether this was due solely to one observation; in that case

$h = k$, $H = n_{ij}$, and $z(h) = y_{ijk}$, i and j both being fixed;

b) to the cell averages for a given level j in table C, when in that case

$h = i$, $H = p_j$, and $z(h) = \bar{y}_{ij}$, j being fixed;

c) to the cell differences, $d_{ij} = y_{ijA} - y_{ijB}$, for a given level of a split-level experiment (table B), when in that case,

$h = i$, $H = p_j$, and $z(h) = d_{ij}$, j being fixed.

13.3 If Dixon's test reveals one of the extreme values in a series (the highest or the lowest) as a straggler or statistical outlier, the test should again be applied to the remaining $H - 1$ values; and if this once more proves one of the extremes as suspect, the test should be applied afresh to the remaining set of $H - 2$ values.

13.4 Again, however, as explained in 12.7, great caution should be exercised in drawing conclusions from the result of repeated applications of Dixon's test. If several stragglers and/or statistical outliers are found at only a single level, this may not be really significant, but if they occur at differing levels within a single laboratory, this may be considered as indicating that that is an outlying laboratory.

13.5 The strategy in dealing with stragglers and/or statistical outliers outlined in 11.6.1 should also be adhered to in the case of Dixon's test.

Table 2 — Critical values for Dixon's outlier test[1)]

Test criterion[2)]	H	Critical values 5 %	Critical values 1 %
$Q_{10} = \dfrac{z(2) - z(1)}{z(H) - z(1)}$ or $\dfrac{z(H) - z(H-1)}{z(H) - z(1)}$ whichever is the greater	3	0,970	0,994
	4	0,829	0,926
	5	0,710	0,821
	6	0,628	0,740
	7	0,569	0,680
$Q_{11} = \dfrac{z(2) - z(1)}{z(H-1) - z(1)}$ or $\dfrac{z(H) - z(H-1)}{z(H) - z(2)}$ whichever is the greater	8	0,608	0,717
	9	0,504	0,672
	10	0,530	0,635
	11	0,502	0,605
	12	0,479	0,579
$Q_{22} = \dfrac{z(3) - z(1)}{z(H-2) - z(1)}$ or $\dfrac{z(H) - z(H-2)}{z(H) - z(3)}$ whichever is the greater	13	0,611	0,697
	14	0,586	0,670
	15	0,565	0,647
	16	0,546	0,627
	17	0,529	0,610
	18	0,514	0,594
	19	0,501	0,580
	20	0,489	0,567
	21	0,478	0,555
	22	0,468	0,544
	23	0,459	0,535
	24	0,451	0,526
	25	0,443	0,517
	26	0,436	0,510
	27	0,429	0,502
	28	0,423	0,495
	29	0,417	0,489
	30	0,412	0,483
	31	0,407	0,477
	32	0,402	0,472
	33	0,397	0,467
	34	0,393	0,462
	35	0,388	0,458
	36	0,384	0,454
	37	0,381	0,450
	38	0,377	0,446
	39	0,374	0,442
	40	0,371	0,438

1) This is R. S. Gardner's version of Dixon's test as published in table 16.[3] This version applies when it is not known at which end of a series of data an outlier may occur.

2) $z(h)$, $h = 1, 2, \ldots, H$, is the series of data to be tested arranged in order of magnitude. The meaning of h, and H in different situations is explained in 13.2.

14 Computation of the mean level m, the repeatability r, and the reproducibility R

14.1 In this International Standard, the method of analysis adopted involves carrying out the computation of m, r and R for each level separately. When there are q levels in all, the results of the computation will be denoted as

m_j, r_j, R_j

$(j = 1, 2, ..., q)$

Subsequently, it is investigated whether r and/or R depend on m, and if so, what is the functional relationship.

14.2 The basic data needed for the computations are presented in three tables (see pages 12 and 13) :

— table A containing the original test results;

— table B containing the measures of within-cell spread, and;

— table C containing the cell averages.

The construction of these tables is explained in 11.9.

14.3 As a consequence of the ruling of 16.9, the number of non-empty cells to be used in the computation will, for a specified level, always be the same in tables B and C. An exception might occur if, owing to missing data, a cell in table A contains only a single test result, which will entail an empty cell in table B but not in table C. In that case, it is possible either

a) to discard the solitary test result which will lead to an empty cell in both tables B and C, or

b) if this is considered an unwarranted loss of information, to insert a nominal value, zero (0), in table B.

Where option b) is taken, the computations have to be carried out in accordance with 14.10, and for a cell with a single test result, any value can be introduced in table B without influencing the final outcome; a nominal zero seems most appropriate.

The number of non-empty cells may be different for different levels; hence the index j in p_j.

14.4 Owing to the rejection of some of the original test results or to missing data, the number of replicates per cell in table A (see 11.9.2) need not necessarily all be the same. This number would therefore be denoted by n_{ij} for laboratory i and level j.

14.5 The computations described below assume that the instructions for rounding specified in 11.9.3 and 11.9.4 have been observed. No further rounding should be carried out in the course of the computations, but an appropriate rounding should be applied to the final results m, r and R.

14.6 The computations can often be simplified, and the risk of computational errors reduced, by coding the data, in particular, using

$$x_{ijk} = b_j (y_{ijk} - a_j)$$

whence $\bar{x}_{ij} = b (\bar{y}_{ij} - a_j)$, $w_{cij} = b_j w_{ij}$, $s_{cij} = b_j s_{ij}$, and $d_{cij} = b_j d_{ij}$, instead of y_{ijk}, $\bar{y}_{ij}$, w_{ij}, s_{ij} and d_{ij}.

a_j is some arbitrary constant chosen to reduce the number of digits to be handled, and b_j some power of 10 chosen so that all or most of the basic data are converted to integers.

It should be noted that, except in the trivial case $b_j = 1$, the values of b_j must be taken into account when computing the final values of s_r^2 and s_L^2 from which r and R are derived.

The procedure is fully illustrated in the example 14.11.3.

14.7 The computational procedure depends on the type of experiment and on the number of replicates in the cells. Four different situations are dealt with in 14.8 to 14.11, each illustrated by a numerical example. They cover most situations likely to arise. In each example, only one level is considered and for convenience the index j has been dropped. In these examples, any outliers found have already been discarded and only the acceptable data are quoted.

14.8 Uniform-level experiment with $n = 2$ replicates per cell

14.8.1 Basic data from tables B and C

Laboratory	Original data	
i	w_i	$\overline{y}_i$
1	0,5	31,45
2	0,0	30,90
3	0,2	30,80
4	0,4	31,30
5	0,3	31,45
6	0,2	31,50
7	0,0	31,40

14.8.2 Computational formulae and numerical results

Number of laboratories : p	$p = 7$
Number of replicates : n	$n = 2$
$S_1 = \Sigma \overline{y}_i$	$S_1 = 218{,}80$
$S_2 = \Sigma \overline{y}_i^2$	$S_2 = 6\ 839{,}555\ 0$
$S_3 = \Sigma w_i^2$	$S_3 = 0{,}58$
$s_r^2 = \dfrac{S_3}{2p}$	$s_r^2 = \dfrac{0{,}58}{2 \times 7} = 0{,}041\ 4$
$*s_L^2 = \left[\dfrac{pS_2 - S_1^2}{p(p-1)}\right] - \dfrac{s_r^2}{2}$	$s_L^2 = \left[\dfrac{7 \times 6\ 839{,}555\ 0 - 218{,}80^2}{7 \times 6}\right] - \dfrac{0{,}041\ 4}{2} = 0{,}061\ 3$
$m = \dfrac{S_1}{p}$	$m = \dfrac{218{,}80}{7} = 31{,}26$
$r = 2{,}83\sqrt{s_r^2}$	$r = 2{,}83\sqrt{0{,}041\ 4} = 0{,}58$
$R = 2{,}83\sqrt{s_L^2 + s_r^2}$	$R = 2{,}83\sqrt{0{,}061\ 3 + 0{,}041\ 4} = 0{,}91$

* If s_L^2 is negative, substitute $s_L^2 = 0$ in the expression of R, to give $R = r$.

14.9 Uniform-level experiment with $n > 2$ replicates per cell

14.9.1 Basic data ($n = 3$)

Laboratory	Original data	
i	s_i	$\bar{y}_i$
1	0,82	28,03
2	1,50	21,25
3	3,00	22,47
4	0,58	25,50
5	1,49	33,08
6	0,50	24,23
7	2,38	20,53
8	0,93	30,17
9	1,07	22,40

14.9.2 Computational formulae and numerical results

Number of laboratories : p	$p = 9$
Number of replicates : n	$n = 3$
$S_1 = \Sigma \bar{y}_i$	$S_1 = 227{,}66$
$S_2 = \Sigma \bar{y}_i^2$	$S_2 = 5\,907{,}243\,4$
$S_3 = \Sigma s_i^2$	$S_3 = 22{,}403\,1$
$s_r^2 = \dfrac{S_3}{p}$	$s_r^2 = \dfrac{22{,}403\,1}{9} = 2{,}489\,2$
$^*s_L^2 = \left[\dfrac{pS_2 - S_1^2}{p(p-1)}\right] - \dfrac{s_r^2}{n}$	$s_L^2 = \left[\dfrac{9 \times 5\,907{,}243\,4 - 227{,}66^2}{9 \times 8}\right] - \dfrac{2{,}489\,2}{3} = 17{,}727\,4$
$m = \dfrac{S_1}{p}$	$m = \dfrac{227{,}66}{9} = 25{,}30$
$r = 2{,}83 \sqrt{s_r^2}$	$r = 2{,}83 \sqrt{2{,}489\,2} = 4{,}46$
$R = 2{,}83 \sqrt{s_L^2 + s_r^2}$	$R = 2{,}83 \sqrt{17{,}727\,4 + 2{,}489\,2} = 12{,}72$

* If s_L^2 is negative, substitute $s_L^2 = 0$ in the expression of R, to give $R = r$.

14.10 Uniform-level experiment with unequal numbers of replicates per cell

14.10.1 Basic data

Laboratory	Original data		Number of replicates
i	s_i	$\bar{y}_i$	n_i
1	0,14	21,30	2
2	0,14	21,50	2
3	0,07	20,75	2
4	0,21	21,75	2
5	0,10	20,90	3
6	0,21	21,05	2
7	0,28	21,50	4
8	0,21	20,85	2
9	0,28	21,10	2
10	0,35	20,85	2
11	(0)	21,30	1

14.10.2 Computational formulae and numerical results

Number of laboratories : p	$p = 11$
$S_1 = \Sigma n_i \bar{y}_i$	$S_1 = 508{,}30$
$S_2 = \Sigma n_i \bar{y}_i^2$	$S_2 = 10\,767{,}765\,0$
$S_3 = \Sigma n_i$	$S_3 = 24$
$S_4 = \Sigma n_i^2$	$S_4 = 58$
$S_5 = \Sigma (n_i - 1) \cdot s_i^2$	$S_5 = 0{,}632\,5$
$s_r^2 = \dfrac{S_5}{S_3 - p}$	$s_r^2 = \dfrac{0{,}632\,5}{24 - 11} = 0{,}048\,6$
$^* s_L^2 = \left[\dfrac{S_2 S_3 - S_1^2}{S_3 (p-1)} - s_r^2\right]\left[\dfrac{S_3 (p-1)}{S_3^2 - S_4}\right]$	$s_L^2 = \left[\dfrac{24 \times 10\,767{,}765\,0 - 580{,}30^2}{24\,(11-1)} - 0{,}048\,6\right]\left[\dfrac{24\,(11-1)}{24^2 - 58}\right] = 0{,}088\,4$
$m = \dfrac{S_1}{S_3}$	$m = \dfrac{508{,}30}{24} = 21{,}18$
$r = 2{,}83 \sqrt{s_r^2}$	$r = 2{,}83 \sqrt{0{,}048\,6} = 0{,}62$
$R = 2{,}83 \sqrt{s_L^2 + s_r^2}$	$R = 2{,}83 \sqrt{0{,}088\,4 + 0{,}048\,6} = 1{,}05$

* If s_L^2 is negative, substitute $s_L^2 = 0$ in the expression of R, to give $R = r$.

14.11 Split-level experiment

14.11.1 Basic data

Laboratory i	Original data d_i	$\overline{y}_i$
1	−0,54	18,770
2	−0,47	18,615
3	−0,43	18,465
4	−0,48	19,660
5	−0,51	18,865
6	−0,49	18,335
7	−0,53	18,895
8	−0,50	18,680
9	−0,57	19,105

14.11.2 Computational formulae and numerical results

Number of laboratories : p $S_1 = \Sigma \overline{y}_i$ $S_2 = \Sigma \overline{y}_i^2$ $S_3 = \Sigma d_i$ $S_4 = \Sigma d_i^2$	$p = 9$ $S_1 = 169{,}390$ $S_2 = 3\,189{,}2$ $S_3 = -4{,}52$ $S_4 = 2{,}283\,8$
$s_r^2 = \dfrac{pS_4 - S_3^2}{2p\,(p-1)}$ $^*s_L^2 = \left[\dfrac{pS_2 - S_1^2}{p\,(p-1)}\right] - \dfrac{s_r^2}{2}$	$s_r^2 = \dfrac{9 \times 2{,}283\,8 - (-4{,}52)^2}{18 \times 8} = 0{,}000\,860$ $s_L^2 = \left[\dfrac{9 \times 3\,189\,2 - 169{,}390^2}{9 \times 8}\right] - \dfrac{0{,}000\,860}{2}$ $= 0{,}152\,050$
$m = \dfrac{S_1}{p}$ $r = 2{,}83\sqrt{s_r^2}$ $R = 2{,}83\sqrt{s_L^2 + s_r^2}$	$m = \dfrac{169{,}39}{9} = 18{,}82$ $r = 2{,}83\sqrt{0{,}000\,860} = 0{,}083$ $R = 2{,}83\sqrt{0{,}152\,050 + 0{,}000\,860} = 1{,}107$

* If s_L^2 is negative, substitute $s_L^2 = 0$ in the expression of R, to give $R = r$.

14.11.3 Computations with coding (see sub-clause 14.6)

14.11.3.1 Basic and coded data

Laboratory	Original data		Coded data	
i	d_i	y_i	d_{ci}	x_i
1	− 0,54	18,770	− 54	77,0
2	− 0,47	18,615	− 47	61,5
3	− 0,43	18,465	− 43	46,5
4	− 0,48	19,660	− 48	166,0
5	− 0,51	18,865	− 51	86,5
6	− 0,49	18,335	− 49	33,5
7	− 0,53	18,895	− 53	89,5
8	− 0,50	18,680	− 50	68,0
9	− 0,57	19,105	− 57	110,5

14.11.3.2 Computational formulae with coded data and numerical results

Number of laboratories : p	$p = 9$
Coding constant : a	$a = 18{,}00$
Coding factor : b	$b = 100$
$x_i = b\,(y_i - a)$	
$d_{ci} = bd_i$	
$S_\mathrm{I} = \Sigma x_i$	$S_\mathrm{I} = 739{,}0$
$S_\mathrm{II} = \Sigma x_i^2$	$S_\mathrm{II} = 72\,878{,}50$
$S_\mathrm{III} = \Sigma d_{ci}$	$S_\mathrm{III} = -452$
$S_\mathrm{IV} = \Sigma d_{ci}^2$	$S_\mathrm{IV} = 22\,838$

It can be seen by comparison with the uncoded calculation in 14.11.2 that :

$$S_\mathrm{I} = b\,(S_1 - pa)$$

$$S_\mathrm{II} = b^2S_2 - 2b^2aS_1 + pb^2a^2$$

$$S_\mathrm{III} = bS_3$$

$$S_\mathrm{IV} = b^2S_4$$

The values of s_r^2 and s_L^2 are then calculated in the normal way :

$s_\mathrm{r}^2 = \dfrac{pS_\mathrm{IV} - S_\mathrm{III}^2}{2p\,(p-1)}$	$s_\mathrm{r}^2 = \dfrac{9 \times 22\,838 - (-452)^2}{18 \times 8}$ $= 8{,}60$
$s_\mathrm{L}^2 = \left[\dfrac{pS_\mathrm{II} - S_\mathrm{I}^2}{p\,(p-1)}\right] - \dfrac{s_\mathrm{r}^2}{2}$	$s_\mathrm{L}^2 = \dfrac{9 \times 72\,878{,}50 - 739{,}0^2}{9 \times 8} - \dfrac{8{,}60}{2}$ $= 1\,520{,}5$

Hence, calculating the decoded mean m, and both r and R,

$m = a + \dfrac{S_\mathrm{I}}{bp}$	$m = 18{,}000 + \dfrac{739{,}0}{100 \times 9} = 18{,}82$
$r = \dfrac{2{,}83}{b}\sqrt{s_\mathrm{r}^2}$	$r = \dfrac{2{,}83}{100}\sqrt{8{,}60} = 0{,}083$
$R = \dfrac{2{,}83}{b}\sqrt{s_\mathrm{L}^2 + s_\mathrm{r}^2}$	$R = \dfrac{2{,}83}{100}\sqrt{1\,520{,}5 + 8{,}60}$ $= 1{,}107$

It can be seen that the final results for m, r and R are identical to those found using the basic data without coding.

15 Establishing a functional relation between r (or R) and m

15.1 It cannot always be taken for granted that there exists a regular functional relation between r (or R) and m. In particular, under the circumstances of 4.2.3, where material heterogeneity forms an inseparable part of the variability of the test results, this will only hold good if this heterogeneity is a regular function of m. With solid materials of different composition and coming from different production processes, this is in no way certain. The point should first be decided before the following procedure is applied. Alternatively, separate values of r and R have to be established for each material investigated.

15.2 The reasoning and computational procedures presented below apply to both r and R; they are presented for r only. Only three types of relationship will be considered :

a) a proportionality relation :

$$r = vm \qquad \ldots (10)$$

b) a linear relation :

$$r = u + vm \qquad \ldots (11)$$

c) a logarithmic relation :

$$\log r = c + d \log m \qquad \ldots (12)$$

or its equivalent :

$$r = Cm^d$$

It is expected that in the majority of cases at least one of these formulae will give a satisfactory fit. If not, the statistical expert carrying out the analysis should seek an alternative solution. To avoid confusion, the constants u, v, ..., occurring in these equations may be distinguished by suffixes u_r, v_r, ..., for r, and u_R, v_R, ..., for R, but these have been omitted in this clause to simplify the notations.

15.3 Equations (10) and also (12) when $d > 0$ (general case), will then lead to $r = 0$ for $m = 0$, which may seem unacceptable from an experimental point of view. However, frequently the values of m encountered in practice will have a lower limit larger than zero such that these equations can be used without introducing serious systematic errors.

15.4 For $u = 0$ and $d = 1$, equations (11) and (12) will be identical to equation (10); and when u lies near zero and/or d lies near unity, two or all three of these equations may yield practically equivalent fits. In that case, equation (10) should be preferred because it involves only one parameter and thus permits the simple statement :

> "Two single test results must be considered as suspect when they differ by more than 100 v %."

15.5 If, in a plot of r_j against m_j, or log r_j against log m_j, the set of points is found to lie reasonably close to a straight line, a line drawn by hand may provide a satisfactory solution; but if for some reason a numerical method of fitting is preferred, the procedure of 15.6 is recommended for equations (10) and (11) and that of 15.8 for equation (12).

15.6 From a statistical point of view, the fitting of a straight line is complicated by the fact that both m and r are estimated. But, since the slope v is usually small — of the order of 0,1 or less — errors in m have little influence and the errors in r predominate. Since, moreover, the purpose is to derive values of r for given values of m, a regression of r on m is appropriate.

15.6.1 This should be a weighted regression because, statistically, the standard error of r is proportional to the value of r. With weights W_j (see 15.6.2) for r_j, the computational formulae are

$$S_1 = \sum_j W_j,\ S_2 = \sum_j W_j m_j,\ S_3 = \sum_j W_j m_j^2,$$

$$S_4 = \sum_j W_j r_j, \text{ and } S_5 = \sum_j W_j m_j r_j.$$

Then, for equation (10)

$$v = S_5/S_3$$

and for equation (11),

$$u = \frac{S_3 S_4 - S_2 S_5}{S_1 S_3 - S_2^2}$$

$$v = \frac{S_1 S_5 - S_2 S_4}{S_1 S_3 - S_2^2}$$

15.6.2 The weights W must be proportional to r^{-2}, but since the values of r_j are subject to errors, the same will hold for the weights. To correct for these and reduce the errors in the final equation as far as possible, the following iterative procedure is recommended :

Writing r_{0j} for the original values of r obtained by one of the procedures of clause 14, we begin by applying the above equations for u or v with weights :

$$W_{0j} = r_{0j}^{-2}$$

$$(j = 1, 2, \ldots, q)$$

which results in equations

$$r_1 = v_1 m \text{ or } r_1 = u_1 + v_1 m, \text{ as the case may be.}$$

From these are computed adjusted values of r_j, namely

$$r_{1j} = v_1 m_j \text{ or } r_{1j} = u_1 + v_1 m_j$$

$$(j = 1, 2, \ldots, q)$$

and the computations are then repeated with the adjusted weights $W_{1j} = r_{1j}^{-2}$

giving

$$r_2 = v_2 m \text{ or } r_2 = u_2 + v_2 m$$

The same procedure could now be repeated once again with weights $W_{2j} = r_{1j}^{-2}$ derived from these equations, but this will only lead to unimportant changes. The step from W_{0j} to W_{1j} is effective in eliminating gross errors in the weights, and the equations for r_2 should be considered as the final result.

15.7 The standard error of log r is approximately proportional to $V(r)$, the coefficient of variation of r. Since the standard error of r is proportional to the value of r, the standard error of log r will be independent of r and an unweighted regression of log r on log m is appropriate when equation (12) is considered.

15.8 For equation (12) the computational formulae are

$$S_1 = \sum_j \log m_j,\ S_2 = \sum_j (\log m_j)^2,$$

$$S_3 = \sum_j \log r_j,\ S_4 = \sum_j (\log m_j)(\log r_j)$$

and

$$c = \frac{S_2 S_3 - S_1 S_4}{q S_2 - S_1^2}$$

$$d = \frac{q S_4 - S_1 S_3}{q S_2 - S_1^2}$$

15.9 Examples of fitting equations (10), (11) and (12) to the same set of data are given in 15.9.1, 15.9.2 and 15.9.3.

15.9.1 Example of fitting equation (10) : $r = vm$

m_j	3,94	8,28	14,18	15,59	20,41
r_{0j}	0,261	0,506	0,359	0,953	1,114
W_{0j}	15	3,8	7,7	1,1	0,81
	$r_1 = 0{,}039\,9\ m$				
r_{1j}	0,156	0,329	0,563	0,619	0,810
W_{1j}	41	9,2	3,2	2,6	1,5
	$r_2 = 0{,}053\,6\ m$				
r_{2j}	0,211	0,444	0,760	0,836	1,094
W_{2j}	22	5,1	1,7	1,4	0,84
	$r_3 = 0{,}053\,6\ m = r_2$				

NOTE — The calculation of r_{2j} shall be done with m_j, r_{0j} and W_{1j}.

15.9.2 Example of fitting equation (11) : $r = u + vm$

	m_j, r_{0j} and W_{0j} as in 15.9.1				
	$r_1 = 0{,}163 + 0{,}025\,2\ m$				
r_{1j}	0,262	0,372	0,520	0,556	0,677
W_{1j}	15	7,2	3,7	3,2	2,2
	$r_2 = 0{,}086 + 0{,}043\,9\ m$				
r_{2j}	0,259	0,449	0,708	0,770	0,982
W_{2j}	15	4,9	2,0	1,7	1,0
	$r_3 = 0{,}092 + 0{,}043\,3\ m$				
r_{3j}	0,263	0,450	0,706	0,767	0,976
	The difference from r_2 is negligible				

NOTE — The values (in 15.9.1 and 15.9.2) of the weights are not of critical importance. Two significant figures suffice.

15.9.3 Example of fitting equation (12) : $\log r = c + d \log m$

$\log m_j$	+ 0,595	+ 0,918	+ 1,152	+ 1,193	+ 1,310
$\log r_{0j}$	− 0,583	− 0,296	− 0,445	− 0,021	+ 0,047
	$\log r = -1{,}053\,2 + 0{,}767\,8 \log m$				
	or $r = 0{,}088\ m^{0{,}77}$, which yields				
r_j	0,253	0,448	0,678	0,729	0,898

15.10 The data used in 15.9 are taken from the case study of clause 23 and have been used here only to illustrate the numerical procedure. They will be further discussed in clause 23.

16 Statistical analysis as a step-by-step procedure

16.1 Collect all available test results in one table — table A (see 11.9.1 and 11.9.2).

16.1.1 It is recommended that this table be arranged into p rows, indexed $i = 1, \ldots, p$, representing the p laboratories that have contributed data, and q columns, indexed $j = 1, 2, \ldots, q$ representing the q levels in increasing order.

16.1.2 In a uniform-level experiment (see 6.1), the test results within a cell of table A need not be distinguished and may be put in any desired order.

16.1.3 In a split-level experiment (see 6.2), it must be clearly stated which of the two test results belongs to sub-level A and which to sub-level B, and the results must be entered in that specific order.

16.2 Inspect table A for any obvious irregularities; investigate, and if necessary discard, any obviously erroneous data and report to the panel.

16.2.1 It is sometimes immediately evident that the test results of a particular laboratory or in a particular cell lie at a level inconsistent with the other data. Such obviously discordant data should be discarded straight away, but the fact must be reported to the panel for further consideration. (See 17.1.)

16.3 From table A, corrected according to 16.2 when needed, compute table B containing measures of within-cell spread, and table C containing the cell averages. (See 11.9.1, 11.9.3 and 11.9.4.)

16.3.1 When a cell in table A of a uniform-level experiment contains only a single test result, one of the options of 14.3 should be adopted. A single test result in a cell of a split-level experiment must be discarded.

16.4 Inspect tables B and C, level by level, for possible stragglers and/or statistical outliers (see 11.6.1).

Apply the statistical tests of clauses 12 and 13 to all suspect items and mark the stragglers with a single, the statistical outliers with a double, asterisk. If there are no stragglers or statistical outliers, go to 16.10.

16.5 Investigate whether there is, or may be, some technical explanation for the stragglers and/or statistical outliers and, if possible, verify such explanations.

Correct or discard, as may be required, those stragglers and/or statistical outliers that have been satisfactorily explained, and apply corresponding corrections to the other tables.

If there are no stragglers or statistical outliers left that have not been explained, go to 16.10.

16.6 If the distribution of the unexplained stragglers or statistical outliers in tables B or C does not suggest any outlying laboratories (see 11.6.4), go to 16.8.

16.7 If the evidence against the suspected outlying laboratories is considered strong enough to justify the rejection of some or all data from these laboratories, then discard the requisite data and report to the panel.

16.7.1 The decision to reject some or all data from a particular laboratory is the responsibility of the statistical expert carrying out the analysis, but must be reported to the panel for further consideration. (See 17.1.)

16.7.2 A large number of stragglers and/or statistical outliers may indicate a pronounced variance in homogeneity or pronounced differences between laboratories, and thereby cast doubt on the suitability of the test method. This again should be reported to the panel.

16.8 If any stragglers and/or statistical outliers remain that have not been explained or attributed to an outlying laboratory, then discard the statistical outliers but retain the stragglers.

16.9 If in the previous steps any cell in table B has been rejected, the corresponding cell in table C must also be rejected, and vice versa.

16.10 From the entries that have been accepted as correct in tables B and C, compute, by the procedures of clause 14, for each level separately :

the mean level m_j, the repeatability r_j and the reproducibility R_j.

16.11 If only a single level has been used, or if it has been decided that the repeatability and reproducibility should be given for each level separately (see 15.1) and not as functions of the level, go to 16.18.

NOTE — The following steps, 16.12 to 16.16, are applied to r and R separately.

16.12 Plot r_j (or R_j) against m_j and judge from this plot whether r (or R) depends on m or not.

If r (or R) is judged to depend on m, then go to 16.14.

If r (or R) is judged to be independent of m, go to 16.13.

If in doubt, work out both cases and let the panel decide.

16.12.1 There still exists no statistical test appropriate for the problem of 16.12. The technical experts familiar with the test method should have sufficient experience to take a decision.

16.13 Use the averages

$$\frac{1}{q}\Sigma\, r_j = r$$

$$\frac{1}{q}\Sigma\, R_j = R$$

as the final values of the repeatability and reproducibility, and go to 16.18.

16.14 Judge from the plot of 16.12 whether the relationship between r (or R) and m can be represented by a straight line, and if so whether equation (10) [$r = vm$ (or $R = vm$)] or equation (11) [$r = u + vm$ (or $R = u + vm$)] is appropriate. (See 15.2.) Determine the parameters v or alternatively u and v, for r and/or R by the procedure of 15.6.

If a linear relationship is considered acceptable, proceed to 16.16. If not, proceed to 16.15.

16.15 Plot log r_j (or log R_j) against log m_j and judge from this plot whether the relation between log r (or log R) and log m can reasonably be represented by a straight line, namely equation (12) [$\log r = c_r + d_r \log m$ (or $\log R = c_R + d_R \log m$)].
[See 15.2 c).]

If this is considered acceptable, compute the values of c_r, d_r (or c_R, d_R) by the procedure of 15.8.

Then the result may be presented either in the form of the equations given above or as follows :

$$r = C_r\, m^{d_r} \text{ (or } R = C_R\, m^{d_R}\text{)}.$$

If a straight line relation is considered acceptable, proceed to 16.16. If not, go to 16.17.

16.16 If a satisfactory relation has been established according to 16.14 or 16.15, then the final values of r and R are the smoothed values obtained from this relationship for given values of m. Go to 16.18.

16.17 If no satisfactory relation according to 16.14 or 16.15 has been established, the statistical expert should decide whether some other relation between r (or R) and m can be established or, alternatively, that the data are so irregular that the establishment of a functional relation must be considered as impossible. Go to 16.18.

16.18 When the final values of r and R have been established, it is possible to verify that they correspond to a 95 % probability, as required by the definitions of 3.1, by means of the data from which they have been computed. Verification may not be strictly needed, but it may serve as a check on the correctness of the computations, and it may indicate a fundamental discrepancy in the test results. However, the application of an appropriate method is difficult and should be done by a statistician. This type of method has not been dealt with in this International Standard.

16.19 The following diagram indicates in a stepwise fashion the procedure given in 16.1 to 16.18.

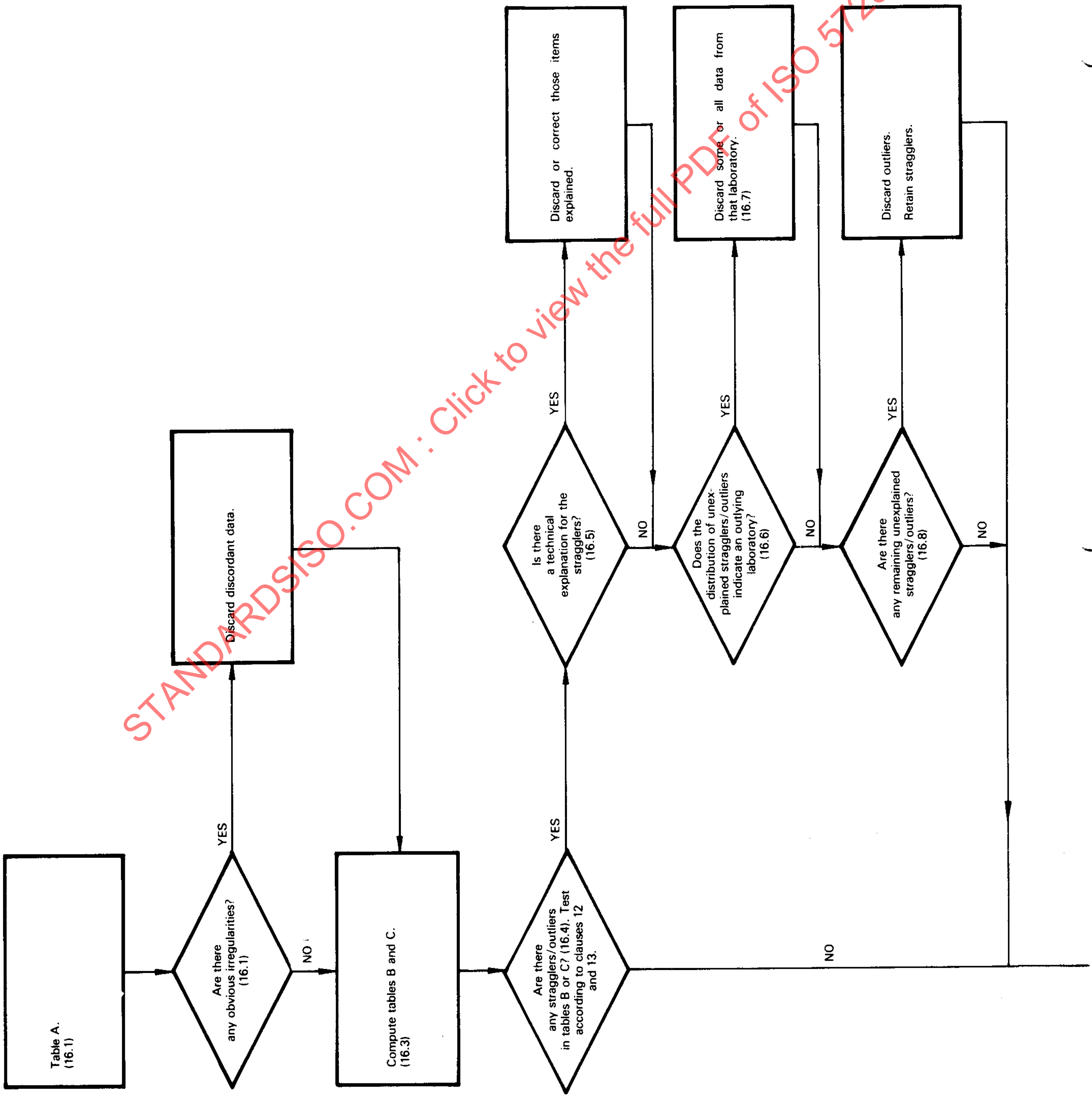
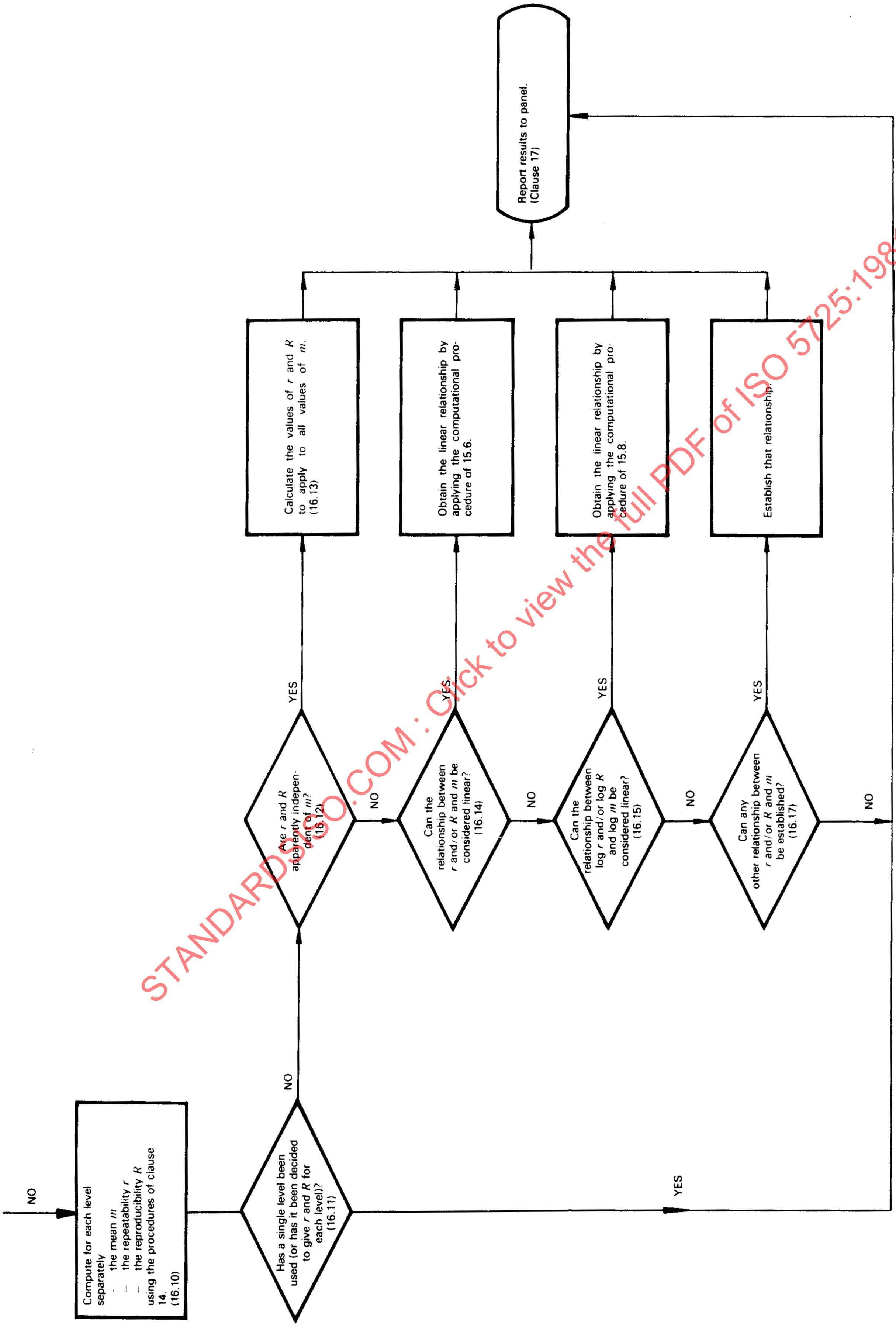


Table A.
(16.1)
Are there
any obvious irregularities?
(16.1)
YES
NO
Discard discordant data.
Compute tables B and C.
(16.3)
Are there
any stragglers/outliers
in tables B or C? (16.4). Test
according to clauses 12
and 13.
YES
NO
Is there
a technical
explanation for the
stragglers?
(16.5)
YES
NO
Discard or correct those items
explained.
Does the
distribution of unex-
plained stragglers/outliers
indicate an outlying
laboratory?
(16.6)
YES
NO
Discard some or all data from
that laboratory.
(16.7)
Are there
any remaining unexplained
stragglers/outliers?
(16.8)
YES
NO
Discard outliers.
Retain stragglers.


NO
Compute for each level separately
– the mean m
– the repeatability r
– the reproducibility R
using the procedures of clause 14.
(16.10)
Has a single level been used (or has it been decided to give r and R for each level)?
(16.11)
YES
NO
Are r and R apparently independent of m?
(16.12)
YES
NO
Calculate the values of r and R to apply to all values of m.
(16.13)
Can the relationship between r and/or R and m be considered linear?
(16.14)
YES
NO
Obtain the linear relationship by applying the computational procedure of 15.6.
Can the relationship between log r and/or log R and log m be considered linear?
(16.15)
YES
NO
Obtain the linear relationship by applying the computational procedure of 15.8.
Can any other relationship between r and/or R and m be established?
(16.17)
YES
NO
Establish that relationship.
Report results to panel.
(Clause 17)

17 Report to, and decisions to be taken by, the panel

17.1 Report

Having completed the statistical analysis, the statistical expert should write a report to be submitted to the panel. In this report, the following information should be given :

a) a full account of the observations received from the operators and/or supervisors concerning the standard for the test method [see 10.6 c)];

b) a full account of the laboratories, if any, that have been rejected as outlying laboratories in steps 16.2 or 16.7, together with the reasons for their rejection;

c) a full account of the stragglers and/or statistical outliers, if any, that were discovered, and whether these were explained, and corrected or discarded;

d) a table of the final results m_j, r_j, and R_j (16.10), and an account of the conclusions reached in steps 16.12, 16.14 or 16.15, illustrated by one of the plots recommended in these steps;

e) tables A, B and C (see 11.9) used in the statistical analysis, possibly as an appendix.

17.2 Decisions

The panel should then discuss this report and take a decision concerning the following questions :

a) Are the discordant test results of rejected outlying laboratories, if any, due to a defect in the description of the standard for the test method? (See 9.1.)

b) What action should be taken with respect to rejected outlying laboratories? (See 17.3.)

c) Do the results of the outlying laboratories and/or the comments received from the operators and supervisors indicate the need to improve the standard for the test method? If so, what are the improvements required?

d) Do the results of the precision experiment justify the establishment of final values of the repeatability and the reproducibility?

e) If so, what are the final values for these parameters, in what form shall they be published, and what is the region in which the precision data apply?

17.3 Outlying laboratories

All laboratories rejected as outliers must be informed of the fact and of the reason for their rejection.

A laboratory rejected on the basis of stragglers and/or outliers in table B will show too high a repeatability variance, which may be due to poor technique or lack of experience of the operator. These laboratories should be encouraged to improve their method, using the established value of the repeatability as a guide (see section four).

A laboratory rejected on the basis of stragglers and/or outliers among the cell averages may be misreading the standard, or be using some instrument with a serious systematic error in its readings. This requires further investigation; the panel should discuss how this can be organized and take corresponding action.

Section four : Utilization of precision data

18 Publication of repeatability and reproducibility

18.1 When a standard test method, for which precision data have been determined, is published, such data shall be included in a section of the method headed "Precision". This section is as much an integral part of the method as other sections on apparatus, reagents, etc.

18.2 The repeatability and reproducibility shall normally be published as a table of three columns giving respectively the range of test results (or a typical result), the repeatability for that range (or level) and the reproducibility for that range (or level) as illustrated below.

Range or level	r	R
From ... to ...		
From ... to ...		
From ... to ...		

18.3 A statement should be added linking the precision to the difference between two results and to the 95 % probability level. A suggested wording is as follows :

"The difference between two single results found on identical test material by one analyst using the same apparatus within a short time-interval will exceed the repeatability on average not more than once in 20 cases in the normal and correct operation of the method.

The difference between two single and independent results found by two operators working in different laboratories on identical test material will exceed the reproducibility on average not more than once in 20 cases in the normal and correct operation of the method."

18.4 A statement can optionally be added that both results shall be considered suspect if the repeatability or reproducibility, as appropriate, is exceeded. Statements regarding subsequent actions, for example, repetition of the test, may also be included in the section on precision.

18.5 In general, a brief mention of the precision experiment should be added to the precision section, possibly as a footnote. A suggested wording is as follows :

"The precision data were determined from an experiment conducted in (year) involving (p) laboratories and (q) levels."

19 Other critical differences derivable from r and R

19.1 The critical differences, as stated in 3.2, are for 95 % probability levels. It is, however, possible to derive the critical differences for other probability levels, and in general, the form will be used that for a critical difference at P % probability is written :

$$\mathrm{Cr\,D}_P\,(\ldots) =$$

19.1.1 Critical difference for probability levels other than 95 %

These can be obtained by multiplying the critical differences for a level of 95 % by the multiplying factors given in the table below.

Multiplying factors for finding critical differences for probability levels other than 95 %

Probability level, P %	Multiplying factor
90	1,645/2 = 0,82
95	2,000/2 = 1,00
98	2,326/2 = 1,16
99	2,576/2 = 1,29
99,5	2,807/2 = 1,40

These multiplying factors result from the assumption that the distribution of the components B and e in the model of 5.1 are normal or approximately normal.

19.2 As defined in 3.1 and 18.3, the uses of r and R are limited to the cases of two single determinations, under either repeatability or reproducibility conditions. It is possible, however, to derive from r and R critical differences for cases other than two single determinations. These are derived below always for the 95 % probability level, and imply that laboratories are effectively selected at random from all the laboratories likely to use the method. For other probability levels, the factors in the table in 19.1.1 can be used.

19.2.1 More than two determinations carried out in one laboratory

If in one laboratory under repeatability conditions two groups of tests are performed, the first group on n_1 tests giving an average of $\bar{y}_1$ and the second group of n_2 tests giving an average of $\bar{y}_2$, then

$$\mathrm{Cr\,D}_{95}\left(|\bar{y}_1 - \bar{y}_2|\right) = r\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}} \qquad \ldots (13)$$

NOTE — If n_1 and n_2 are both unity, this reduces to r, as expected.

19.2.2 Two laboratories each doing more than one determination

If one laboratory performs n_1 determinations with mean value $\bar{y}_1$, while the second laboratory performs n_2 determinations with mean value $\bar{y}_2$, then

$$\mathrm{Cr\,D}_{95}\left(|\bar{y}_1 - \bar{y}_2|\right) = \sqrt{R^2 - r^2\left(1 - \frac{1}{2\,n_1} - \frac{1}{2\,n_2}\right)} \quad \ldots (14)$$

In particular, if $n_1 = n_2 = 1$, this reduces to R as expected, and if $n_1 = n_2 = 2$, this reduces to $\sqrt{R^2 - \frac{r^2}{2}}$

19.2.3 Comparison with a reference level for one laboratory

If n determinations performed by one laboratory under repeatability conditions produce a mean value $\bar{y}$ which is to be compared with a reference value m_o (such as a value specified in a contract), assuming that $m = m_o$, then

$$\mathrm{Cr\,D}_{95}\left(|\bar{y} - m_o|\right) = \frac{1}{\sqrt{2}}\sqrt{R^2 - r^2\left(\frac{n-1}{n}\right)} \quad \ldots (15)$$

19.2.4 Comparison with a reference level for several laboratories

If p laboratories have performed n_i determinations giving averages $\bar{y}_i$ and an overall average

$$\bar{\bar{y}} = \frac{1}{p}\sum \bar{y}_i$$

$(i = 1, 2, \ldots, p)$

the critical difference for comparing this overall average with a reference value m_o, assuming that $m = m_o$, is

$$\mathrm{Cr\,D}_{95}\left(|\bar{\bar{y}} - m_o|\right) = \frac{1}{\sqrt{2}\,p}\sqrt{R^2 - r^2\left(1 - \frac{1}{p}\sum \frac{1}{n_i}\right)} \quad \ldots (16)$$

19.2.5 When in comparing two averages or a single average with a reference value the absolute difference exceeds the corresponding critical difference as given above, then the difference should be considered as suspect; there may be an assignable cause and this should be investigated. In particular, when the reference value in 19.2.4 is a true value, a suspect difference may indicate that the test method has a bias (see 5.2.2).

20 Practical applications

Practical applications of repeatability and reproducibility, including procedures to be followed when a difference between observations exceeds the repeatability or reproducibility criteria, will be the subject of a future International Standard.

Section five : Examples

21 Introduction

21.1 In a report on the results of a precision experiment, full details should be given concerning the standard specifying the test method and the way the samples have been prepared. In most of the examples of this section, that information is missing. In the literature, data are often used to illustrate the statistical analysis, and the test method by which they were obtained is immaterial and not mentioned.

21.2 Thus, the main purpose of the examples of this section is to show how the analysis by the step-by-step procedure of clause 16 works out in practice. In addition, however, these examples show that this systematic analysis does not always tell the complete story. Not infrequently, an attentive statistician will notice peculiarities in the data that are not covered by the tests prescribed in clauses 12 and 13 and this will induce him to apply some further criteria or graphical presentation. It is impossible in a standard like the present one to cover all possible variations and a few examples must suffice. They demonstrate why the analysis should preferably be carried out by a statistician experienced in the analysis of experimental data.

21.3 A brief review of the four case studies presented below follows :

Clause 22 : 16 laboratories, 4 levels, duplicate test results.
There is one empty cell and one missing test result. A tally of the ranges points to a lack of normality which might warrant further investigation.

Clause 23 : 9 laboratories, 5 levels, duplicate test results.
One laboratory is rejected as an outlying laboratory, and the data of one other cell are rejected because the data and additional information indicate that a wrong sample may have been analysed. r and R are strongly dependent on m, and it is difficult to decide which is the best formula to be adopted.

Clause 24 : 8 laboratories, 4 levels, the numbers of test results per cell vary from 3 to 5.
There are no outliers. The results indicate a dependence of r and R on m, but this is not quite certain. The panel should decide on the basis of existing experience.

Clause 25 : A split-level experiment with 25 laboratories and 1 level.
This is an example where there is no problem.

22 Determination of the softening point of pitch

22.1 Background

22.1.1 Source

Standard methods for testing tar and its products. 6th/7th edition. STPTC pitch section (Method PT3 using neutral glycerin).[5]

22.1.2 Material

Selected from commercial batches of pitch collected and prepared as specified on page 501 of the source manual.

22.1.3 Description

This was the determination of a physical property involving temperature measurement in degrees Celsius. Sixteen laboratories co-operated, and it was intended to test four specimens at about 87,5 — 92,5 — 97,5 and 102,5 °C to cover the normal range of products, but the wrong material was chosen for specimen 2 with a mean temperature of about 96 °C, similiar to specimen 3. Laboratory 5 applied the method incorrectly at first on specimen 2 (the first one they tested) and there was then insufficient material left for more than one determination. Laboratory 8 found they did not have specimen 1 (they had specimen 4 in duplicate).

22.2 Inspection of table 22A for suspect entries

There are no obvious stragglers or statistical outliers and no statistical tests are required at this stage.

Table 22A — Original test results (y_{ijk}, °C)

Laboratory i \ Level j	1		2		3		4	
Test k	1	2	1	2	1	2	1	2
1	91,0	89,6	97,0	97,2	96,5	97,0	104,0	104,0
2	89,7	89,8	98,5	97,2	97,2	97,0	102,6	103,6
3	88,0	87,5	97,8	94,5	94,2	95,8	103,0	99,5
4	89,2	88,5	96,8	97,5	96,0	98,0	102,5	103,5
5	89,0	90,0	97,2	—	98,2	98,5	101,0	100,2
6	88,5	90,5	97,8	97,2	99,5	103,2	102,2	102,0
7	88,9	88,2	96,6	97,5	98,2	99,0	102,8	102,2
8	—	—	96,0	97,5	98,4	97,4	102,6	103,9
9	90,1	88,4	95,5	96,8	98,2	96,7	102,8	102,0
10	86,0	85,8	95,2	95,0	94,8	93,0	99,8	100,8
11	87,6	84,4	93,2	93,4	93,6	93,9	98,2	97,8
12	88,2	87,4	95,8	95,4	95,8	95,4	101,7	101,2
13	91,0	90,4	98,2	99,5	98,0	97,0	104,5	105,6
14	87,5	87,8	97,0	95,5	97,1	96,6	105,2	101,8
15	87,5	87,6	95,0	95,2	97,8	99,2	101,5	100,9
16	88,8	85,0	95,0	93,2	97,2	97,8	99,5	99,8

22.3 Computation and inspection of cell variabilities and averages

22.3.1 Cell ranges (table 22B)

In this example there are 2 results per cell and the ranges can be used to represent the variability.

Table 22B — Cell ranges (w_{ij}, °C)

Laboratory i \ Level j	1	2	3	4
1	1,4	0,2	0,5	0,0
2	0,1	1,3	0,2	1,0
3	0,5	3,3	1,6	3,5
4	0,7	0,7	2,0	1,0
5	1,0	—	0,3	0,8
6	2,0	0,6	3,7	0,2
7	0,7	0,9	0,8	0,6
8	—	1,5	1,0	1,3
9	1,7	1,3	1,5	0,8
10	0,2	0,2	1,8	1,0
11	3,2	0,2	0,3	0,4
12	0,8	0,4	0,4	0,5
13	0,6	1,3	1,0	1,1
14	0,3	1,5	0,5	3,4
15	0,1	0,2	1,4	0,6
16	3,8	1,8	0,6	0,3

Application of Cochran's test leads to the following values of the test statistic C :

$j = 1$; $C = 0{,}391$ (15)
$j = 2$; $C = 0{,}424$ (15)
$j = 3$; $C = 0{,}434$ (16)
$j = 4$; $C = 0{,}380$ (16)

The numbers within parentheses indicate the number of ranges included.

Comparing with the critical values of table 1, we find that no value exceeds 0,452, the 5 % critical value for $p = 16$ and $n = 2$, hence no stragglers or statistical outliers are detected.

22.3.2 Cell averages (table 22C)

Table 22C — Cell averages ($\bar{y}_{ij}$, °C)

Laboratory i \ Level j	1	2	3	4
1	90,30	97,10	96,75	104,00
2	89,75	97,85	97,10	103,10
3	87,75	96,15	95,00	101,25
4	88,85	97,15	97,00	103,00
5	89,50	—	98,35	100,60
6	89,50	97,50	101,35	102,10
7	88,55	97,05	98,60	102,50
8	—	96,75	97,90	103,25
9	89,25	96,15	97,45	102,40
10	85,90	95,10	93,90	100,30
11	86,00	93,30	93,75	98,00
12	87,80	95,60	95,60	101,45
13	90,70	98,85	97,50	105,05
14	87,65	96,25	96,85	103,50
15	87,55	95,10	98,50	101,20
16	86,90	94,10	97,50	99,65

NOTE — The entry for $i = 5$, $j = 2$ has been dropped (see 14.3).

Taking $j = 3$ as an example, Dixon's statistic is

$$Q_{22} = \text{the larger of } \frac{95{,}00 - 93{,}75}{98{,}50 - 93{,}75}$$

$$\text{and } \frac{101{,}35 - 98{,}50}{101{,}35 - 95{,}00}$$

so $Q_{22} = 0{,}449$