International Standard 5725

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

Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests

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

Second edition — 1986-09-15

UDC 519.248 : 620.1

Ref. No. ISO 5725-1986 (E)

Descriptors : tests, reproducibility, statistical analysis.

Price based on 49 pages

Foreword

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

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. They are approved in accordance with ISO procedures requiring at least 75 % approval by the member bodies voting.

International Standard ISO 5725 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods.*

This second edition cancels and replaces the first edition (ISO 5725-1981), of which it constitutes a minor revision.

Users should note that all International Standards undergo revision from time to time and that any reference made herein to any other International Standard implies its latest edition, unless otherwise stated.

Printed in Switzerland

Contents

Page

0 Introduction 1
1 Scope 1
2 Field of application 1

Section one : General principles

3 Quantitative definitions of repeatability and reproducibility of a standard test method 3
4 Practical implications of the definitions 4
5 Statistical model 6
6 Design of a precision experiment 7
7 Analysis of the data 7

Section two : Organization of an inter-laboratory precision experiment

8 Personnel requirements 8
9 Tasks and problems 8
10 Comments on clauses 8 and 9 9

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

11 Preliminary considerations 12
12 Cochran's test 15
13 Dixon's test 16
14 Computation of the mean level m, the repeatability r and the reproducibility R 17
15 Establishing a functional relationship between r (or R) and m 24
16 Statistical analysis as a step-by-step procedure 25
17 Reporting to, and decisions to be taken by, the panel 27

Section four : Utilization of precision data

18 Publication of repeatability and reproducibility values 30
19 Other critical differences derivable from r and R 30
20 Practical applications 31

Section five : Examples

21 General information 32
22 Uniform-level experiment with no missing or outlying data 32
23 Uniform-level experiment with missing data 36
24 Uniform-level experiment with outlying data 40
25 Split-level experiment with a single level 44

Annexes

A Critical values for Cochran's test 46
B Critical values for Dixon's test 47
C Symbols and subscripts 48

Bibliography 49

Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests

0 Introduction

0.1 Tests performed on presumably identical material (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 some specified value may be within the scope of unavoidable random errors, in which case a real deviation from such a specified value 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 variations between supposedly identical specimens) may contribute to the variability of a test procedure, including the following :

a) the operator;

b) the equipment used;

c) the calibration of the equipment;

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

The variability between tests performed by different operators and/or with different equipment will usually be greater than between tests carried out by a single operator using the same equipment.

0.3 Precision is a general term for the variability between repeated tests. Two measures of precision, termed repeatability and reproducibility, have been found necessary and, for many practical cases, sufficient for describing the variability of a test method. Repeatability refers to tests performed under conditions that are as constant as possible, with the tests performed during a short interval of time (see 4.3) in one laboratory by one operator using the same equipment. On the other hand, reproducibility refers to tests performed in widely varying conditions, in different laboratories with different operators and different equipment. Under repeatability conditions factors a) to d) listed in 0.2 are considered constants and do not contribute to the variability, while under reproducibility conditions they vary and contribute to the variability of the test results. Thus repeatability and reproducibility are two extremes, the first measuring the minimum and the second the maximum variability in results. Other intermediate measures of variability between these two extremes are conceivable, such as repetition of tests within a laboratory at longer time intervals, or by different operators, or including the effects of recalibration but these are not considered in this International Standard. If, in a particular situation, some intermediate measure should be needed, it must be clearly defined by some responsible authority, together with the circumstances under which it applies and the method by which it should be determined.

0.4 The definitions used in this International Standard are given in clause 3 and the symbols and subscripts used are given in annex C.

A bibliography of the publications referred to in this International Standard is appended.

1 Scope

This International Standard establishes practical definitions of repeatability r and reproducibility R which lend themselves to numerical estimation by experiment (see clause 3). It does not provide any measure of the errors in estimating the values of r and R. It discusses the implications of these definitions of r and R.

This International Standard establishes basic principles for the layout, organization and analysis of experiments designed for estimating r and R (see clauses 6 to 17). Experiments for this purpose will be referred to as precision experiments. Only the simplest type of experiment for the estimation of r and R is considered, which consists of tests on samples of identical material by several laboratories.

This International Standard also presents rules for the interpretation and application of these estimates of r and R in practical situations (see clauses 18 to 20).

This International Standard does not deal with determining the accuracy of the test method, as measured by the difference between the overall mean value and the true value or conventional true value.

2 Field of application

This International Standard deals exclusively with test methods which yield a single numerical figure as the test result, although this single figure may be the outcome of a calculation from a set of observations.

The essence of the determination of precision values is that they measure the ability of a test method to repeat a given determination. Thus the implication is that exactly the same thing is being measured in exactly the same way.

In order that the measurements are made in the same way, the test method shall have been standardized and in use in a number of different laboratories. All tests forming part of a precision experiment shall be carried out in accordance with that standard.

Ideally, the various tests should be carried out using the same specimen. Unfortunately many tests are destructive in nature (chemical tests, strength tests of materials) so that the same specimen is not available for further determinations. Under such circumstances, different specimens shall be used, but to conform to the basic principle every effort shall be made to ensure that the specimens are as nearly identical as is possible. Care shall also be taken that the specimens are not just identical when the samples are prepared, but that they should also be identical at the time of testing.

Because of the above principles, precision should not be determined using specimens which are known not to be, or are even suspected of not being identical. Thus the specimens for test should be taken as similar sub-samples of one bulk sample, and shall never be drawn from different lots or different consignments. These points are discussed further in 4.2.

In practice, where destructive testing is involved, the contribution to the variability in the test results arising from differences between the specimens on which the tests are performed shall either be negligible compared to the variability of the test method itself or else form an inherent part of the test method, and thus truly a component of precision (see 4.2).

The statistical model described in clause 5 is accepted as a suitable basis for the interpretation and analysis of the test results given by a precision experiment which conforms to the principles stated above.

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; the following definitions conform with ISO 3534[1].

3.1.1 observed value : The value of a characteristic determined as the result of an observation.

NOTE — This is a single value obtained from a single observation.

3.1.2 test result : The value of a characteristic determined by carrying out a specified test method.

NOTE — The test method may specify that a number of individual observations be made, and their average reported as the test result. It may also require standard corrections to be applied, such as correction of gas volumes to standard temperature and pressure. Thus a single test result can be a result calculated from several observed values.

3.1.3 level of the test : The general average of the test results from all laboratories for one particular material or specimen tested.

3.1.4 cell : The test results at a single level obtained by one laboratory.

3.1.5 precision : The closeness of agreement between mutually independent test results obtained under stipulated conditions.

NOTES

1 Precision depends only on the distribution of random errors and does not relate to the true value, conventional true value or specified value.

2 Repeatability and reproducibility are concepts of precision.

3.1.6 repeatability : The closeness of agreement between mutually independent test results obtained under repeatability conditions.

3.1.7 repeatability conditions : Conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time.

3.1.8 repeatability standard deviation : The standard deviation of test results obtained under repeatability conditions. It is a parameter of dispersion of the distribution of test results under repeatability conditions.

NOTE — Similarly, repeatability variance and repeatability coefficient of variation could be defined and used as parameters of dispersion of test results under repeatability conditions.

3.1.9 repeatability value, r : The value below which the absolute difference between two single test results obtained under repeatability conditions may be expected to lie with a probability of 95 %.

NOTE — For brevity, in the remainder of this International Standard, "repeatability value" is shortened to "repeatability" or just "r" where the context makes it clear that it is the values that are referred to.

3.1.10 repeatability critical difference : The value below which the absolute difference between two single test results obtained under repeatability conditions may be expected to lie with a specified probability.

NOTES

1 The specified probability has to be attached as a subscript to the symbol r of the repeatability critical difference, for example r_{90} is the repeatability critical difference for a probability of 90 %.

2 The repeatability value r is the repeatability critical difference for a probability of 95 %, the subscript being omitted in this special case.

3.1.11 reproducibility : The closeness of agreement between test results obtained under reproducibility conditions.

3.1.12 reproducibility conditions : Conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment.

3.1.13 reproducibility standard deviation : The standard deviation of test results obtained under reproducibility conditions. It is a parameter of dispersion of the distribution of test results under reproducibility conditions.

NOTE — Similarly, reproducibility variance and reproducibility coefficient of variation could be defined and used as parameters of dispersion of test results under reproducibility conditions.

3.1.14 reproducibility value, R : The value below which the absolute difference between two single test results obtained under reproducibility conditions may be expected to lie with a probability of 95 %.

NOTE — For brevity, in the remainder of this International Standard, "reproducibility value" is shortened to "reproducibility" or just "R" where the context makes it clear that it is the values that are referred to.

3.1.15 reproducibility critical difference : The value below which the absolute difference between two single test results obtained under reproducibility conditions may be expected to lie with a specified probability.

NOTES

1 The specified probability has to be attached as a subscript to the symbol R of the reproducibility critical difference, for example R_{90} is the reproducibility critical difference for a probability of 90 %.

2 The reproducibility value R is the reproducibility critical difference for a probability of 95 %, the subscript being omitted in this special case.

3.2 The definitions given in 3.1 apply to results variable on a continuous scale. If the test result is discrete or 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 probability of not less than the specified value.

3.3 The terms "repeatability" and "reproducibility" have been adopted because they have been in common use for several years. The symbols r and R are already in general use for other purposes; 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 if the full wordings "repeatability r" and "reproducibility R" are used whenever there is a possibility of misunderstanding; particularly when r and R are quoted in standards.

3.4 R and r as defined in this International Standard are meant in the first place as criteria by which to judge how far a difference between two single test results can be ascribed to random fluctuations; a difference larger than r or R is suspect and may justify the conclusion that there exists a systematic difference, or lead to some additional investigation. In this sense, r and R can be termed critical differences, to be applied to a pair of test results obtained under repeatability and reproducibility conditions respectively.

3.4.1 It is sometimes required to compare the averages of two or more tests or to compare the average of a series with a specified value. Critical differences for such purposes can be derived from r and R as explained in 19.2.1 to 19.2.4.

3.4.2 As defined, r and R are associated with a probability level of 95 %. Sometimes critical differences with a probability level other than 95 % may be preferred; these can be computed as explained in 19.1.1. In such cases, to avoid misinterpretations, the probability level should then be attached as a subscript; for example r_{99} or R_{90}.

3.4.3 The definitions in 3.1.9 and 3.1.14 refer to theoretical constants which in reality remain unknown. The values of r and R actually determined from a precision experiment as described in this International Standard are, in statistical terms, estimates of these constants, and as such are subject to errors. Consequently, the probability levels associated with r and R will not be exactly 95 % but only of the order of 95 %, and this will also be true for other critical differences derived from them. This is unavoidable but does not seriously detract from their practical value as they are primarily designed to serve as tools for judging whether the difference between results could be ascribed to random uncertainties inherent in the test method or not. Differences larger than r or R are suspect.

3.5 If the requirements of this International Standard concerning the number of laboratories to be included in a precision experiment and the number of tests they should each carry out are followed (see 10.1), the resulting estimates of r and R will be sufficiently precise for practical purposes, with the proviso that the laboratories participating are truly representative of all laboratories using the standard method. This hypothetical population is defined by requirements similar to those given in 10.2. If at some future date it should become evident that this condition was not or is no longer satisfied by the original precision experiment, then a fresh precision experiment may be required, unless it should be possible to re-estimate r and R to conform to the altered conditions.

3.6 In principle, repeatability r, as defined in 3.1.9, can be applied to any test method within any single laboratory. A basic assumption underlying this International Standard is that, for a standardized test method, repeatability will be, at least approximately, the same for all laboratories applying the standard method, so that it is possible to establish one common average repeatability applicable to any laboratory. However, any laboratory can, by carrying out a series of tests under repeatability conditions, arrive at an estimate of its own particular repeatability for the test method, and check it against the common standard value. Such a procedure has not been worked out in detail in this International Standard.

3.7 When the reproducibility is to be used as a critical difference, the pair of test results to be compared shall have been obtained from two laboratories selected at random from the total population of laboratories using the standard test method. Where test results are always compared between the same two laboratories, caution is needed, because the probability level associated with R may then no longer hold true owing to a possible systematic difference between the results from these two particular laboratories. If it is thought that this may be the case, the two laboratories in question should organize a precision experiment between themselves in order to determine the magnitude of this systematic difference.

3.8 Although throughout this International Standard repeatability and reproducibility are considered in terms of critical differences, there is no reason for preventing the expression of precision results in terms of standard deviations or coefficients of variation if, for any particular application, this would be more appropriate.

3.9 The values of r and R, once determined, can be used in a variety of ways. For example, they can serve

— to verify that the experimental technique of a laboratory is up to standard (see 3.6);

— in designing quality control procedures;

— in comparing test results from a batch of material with a product specification;

— in designing the specifications in the first place to ensure that conformity is verifiable by the test method;

— in comparing test results on the same batch of material obtained by a supplier and a consumer;

— in assessing the suitability of rival test methods.

In some applications, various other factors may have to be taken into consideration, for example see 4.2.6.

4 Practical implications of the definitions

4.1 Standard test method

4.1.1 As stated in clause 2, the test method under investigation has to be one that has been standardized. This means that

there has to be a standard, i.e. a written document that lays down in full detail how the test should be carried out, preferably including a description as to how the test specimen should be obtained and prepared. The estimates of r and R derived from such an experiment should always be quoted as valid only for tests carried out in accordance with the standard method.

4.1.2 The existence of a standard for the test method implies the existence of an organization responsible for the establishment of the standard under study.

4.1.3 Preparing a standard for a test method requires a careful evaluation of the method (or possibly several alternative methods) by means of experiments in which a number of laboratories take part. Such a standardization experiment will provide some preliminary information concerning the values of r and R. The essential points underlying a precision experiment to determine r and R is that it will usually require the cooperation of a larger number of laboratories than for a standardization experiment, and that these laboratories shall be recruited from among all those using, or likely to use, the standard in normal operations and not exclusively consist of laboratories that have gained special experience during the process of standardizing the method. Thus a precision experiment arranged for the determination of r and R should not as a rule be organized until after the standard for the test method has been issued and is in general use. This does not mean, however, that any information regarding the possible values of r and R obtained from a standardization experiment is of no value, as they can be taken into consideration when designing the precision experiment.

4.1.4 A precision experiment can also be considered as a practical test of the adequacy of the standard. One of the main purposes of standardization is to eliminate differences between users (laboratories) as far as possible, and the data provided by a precision experiment will reveal how far this purpose has been achieved. Pronounced differences 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 c), 17.2 b) and c) and 17.3.]

4.2 Identical material

4.2.1 In a precision experiment, samples of a specific material or specimens of a specific product are sent from a central point to a number of laboratories in different places, different countries, or even in different continents. The requirement that the tests in these laboratories shall be made on identical material refers to the moment when these tests are actually carried out, and in order to achieve this the following two different conditions have to be satisfied :

a) the samples have to be identical when despatched to the laboratories, and

b) they have to remain identical during transport and during the different time intervals that may elapse before the tests are actually performed in the participating laboratories.

In organizing precision experiments, both conditions shall be carefully observed.

4.2.2 A fluid or fine powder can be homogenized by stirring, and samples drawn from such batches can then be considered as identical at the moment they are prepared. Additional precautions may be needed to ensure that they remain identical up to the time the tests are carried out. 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 example during transport. When reaction with the atmosphere may be expected, the specimens may be sealed into ampoules, either evacuated or filled with an inert gas. For perishable materials, such as foodstuffs or blood samples, it may be necessary to send them to the participating laboratories in a deep-frozen state with detailed instructions of the procedure for thawing. Each case has to be judged on its merits.

4.2.3 When the tests have to be performed on discrete objects that are not altered by testing, they 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.

4.2.4 When tests have to be performed on solid materials that cannot be homogenized (such as metals, rubber or textile fabrics) and when the tests cannot be repeated on the same test piece, inhomogeneity in the test material will form an essential component of the precision of the measurement and the idea of identical material no longer holds good. Precision experiments can still be carried out, but the values of r and R may only be valid for the particular material used and should be quoted as such. A more universal use of r and R will be acceptable only if it can be demonstrated that the values do not differ significantly between material produced at different times or by different producers. This would require a more elaborate experiment than has been considered in this International Standard.

4.2.5 In 4.2.1 to 4.2.4, reference is made to testing in different laboratories, with the implication of transportation of the test specimens to the laboratory, but some test specimens are not transportable, such as an oil storage tank. In such cases, testing by different laboratories means that different operators are sent with their equipment to the test site. In other cases, the quantity being measured may be transitory or variable, such as water flow in a river, when care shall be taken that the different measurements are made under as near as possible the same conditions. The guiding principle shall always be that the objective is to determine the ability to repeat the same measurement.

4.2.6 In practice, r and R, or other critical differences derived from them using the methods specified in 19.1.1 and/or 19.1.2, are often used in order to compare batches of commercial material with a specification or to compare two batches with each other. A difference larger than that critical difference can then, *inter alia*, be explained by the normal commercial inhomogeneity in the batches of material unless it has been possible to include this lack of homogeneity in the determination of r and R. However, in that case, the difficulties will be the same as those mentioned in 4.2.4.

4.3 Short intervals of time

According to the definition in 3.1.7, tests for the determination of repeatability have to be made under constant operating conditions, i.e. during the time covered by the tests, factors such as those in 0.2 should be constant. In particular, the equipment should not be recalibrated between the tests unless this is an essential part of every single determination. In practice, tests under repeatability conditions should be conducted in as short a time as possible in order to minimize changes in those factors, such as environmental ones, which cannot always be guaranteed constant. [See 10.4.1 c).]

5 Statistical model

5.1 Basic model

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 general average;

B is the between-laboratory variation;

e is the random error occurring in every test.

Other models are sometimes used, but the above will cover the majority of practical cases. (See 5.6.)

5.2 General average, m

5.2.1 The general average, m, of the material tested is called the "level of the test property"; specimens of different purities of a chemical or different materials (e.g. different types of steel) will correspond to different levels. In many technical situations, the level of the test property is exclusively defined by the test method, and the notion of an independent true value does not apply. However, in some situations, the concept of a true value μ of the test property may hold good, such as the true concentration of a solution that is being titrated. The level m is not necessarily equal to the true value μ; the difference $(m - \mu)$, when it exists, is called the "bias of the test method".

5.2.2 When r and R are used to test the difference between test results, a bias will have no influence and can be ignored. But when these criteria 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 where the contract or specification refers to the true value, μ, and not to the test level, m. If a true value exists and is known, the analysis of a precision experiment can indicate that there is a bias. (See note to 19.2.4.)

5.3 Term B in the basic model (see 5.1)

5.3.1 The term B is considered to be constant during any series of tests performed under repeatability conditions, but it is considered to behave as a random variable in a series of tests performed under reproducibility conditions. The procedures given in this International Standard were developed assuming that the distribution of this error variable was approximately normal, but, in practice, they work for most distributions provided that they are unimodal and that the critical differences are for the 95 % level. Its variance is called the between-laboratory variance and is expressed as

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

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

5.3.2 In general, B can be considered as the sum of both random and systematic components, but they are not separated in this analysis. No attempt has been made in this International Standard to give an exhaustive list of the factors that contribute to B, but they include different climatic conditions, variations of equipment within the manufacturer's tolerances, and even the techniques in which operators are trained in different places.

5.3.3 If there are a limited number of laboratories likely to use the method at any time, B can only take a limited number of values, and to be of practical use, reproducibility shall be determined from a set of laboratories which can be considered as selected at random from all those likely to use the method. Some caution is needed when the test results to be compared are always performed by the same laboratories. An example of the sort of problem that can arise in this situation is given in clause 23, in which the results from two (11 and 1) of the laboratories are shown to differ consistently by about 4 °C. Where only two laboratories are regularly concerned, reproducibility as such should not be used, but a cooperative experiment between the two laboratories to determine their relative bias, and thus their own specific reproducibility, should be carried out.

5.4 Error term e in the basic model (see 5.1)

5.4.1 The term e represents a random error occurring in every single test result and the procedures given in this International Standard were developed assuming that the distribution of this error variable was approximately normal, but, in practice, they work for most distributions provided that they are unimodal and that the critical differences are for the 95 % level. Within a single laboratory its variance is called the within-laboratory variance and is expressed as

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

5.4.2 It may be expected that σ_w^2 will have different values in different laboratories due to differences such as in the skills of the operators, but, in this International Standard, it is assumed that for a properly standardized test method such differences between laboratories should be small and that it is justifiable to establish a common value of within-laboratory variance for all the laboratories using the test method. This common value, which is the average of all the within-laboratory variances taken over all the labaoratories taking part in the precision experiment, is called the repeatability variance and is expressed as

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

5.5 Relation between the basic model, and r and R

When the basic model (see 5.1) is adopted, the repeatability value r depends solely on the repeatability variance (5.4.2), while the reproducibility value R depends on the sum of the repeatability variance and the between-laboratory variance (see 5.3.1). Thus, there are two quantities, called the repeatability standard deviation, expressed as

$$\sigma_r = \sqrt{\text{var}(e)}$$

and the reproducibility standard deviation, expressed as

$$\sigma_R = \sqrt{\sigma_L^2 + \sigma_r^2}$$

Hence

repeatability value $r = f\sqrt{2}\,\sigma_r$, and ...(2)

reproducibility value $R = f\sqrt{2}\,\sigma_R$ ...(3)

where

the coefficient $\sqrt{2}$ is derived from the fact that r and R refer to the difference between two single test results;

f is a factor the value of which depends both on the number of test results available for estimating each of the variances and on the shape of the distributions of the components B and e (see 5.1).

However, if these distributions are approximately normal and the number of test results is not too small, then for a probability level of 95 % the factor f will never differ much from the value 2 and the use of this value is therefore recommended in this International Standard, with the value of $f\sqrt{2}$ being rounded to be 2,8. (Taking into account variations in the factor f would lead to considerable complications and would not effectively contribute to the practical value of r and R.)

In practice, as the exact values of the repeatability standard deviation and the reproducibility standard deviation are not known, they are replaced by their estimates s, leading to

$$r = 2{,}8\,s_r \quad ...(4)$$

$$R = 2{,}8\,s_R \quad ...(5)$$

5.6 Suitability of the basic model

It is clear that the basic 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. For the purposes of this International Standard, the model is an acceptable approximation as long as the experimental requirements laid down in section two are met and the statistical tests specified in section three do not yield significant results that indicate its unsuitability. The action that should be taken when these statistical tests indicate that the model is unsuitable are discussed in clauses 16 and 17.

6 Design of a precision experiment

6.1 In one layout, samples from q batches of material, representing q different levels of the test property, are sent to p laboratories which each perform n tests under repeatability conditions at each level. These n tests are thus made on identical material. This type of experiment is 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 Full examples of both layouts are given in the case studies in section five. Practical considerations in planning and execution are given in section two.

7 Analysis of the data

7.1 The analysis of the data produced by a precision experiment, which should be considered as a statistical problem to be entrusted to a statistical expert (see 8.2 and 9.2), involves the following three successive stages :

a) critical examination of the data in order to identify and treat outliers or other irregularities and 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 either of the first two depend on the level m.

7.2 As detailed in 14.7 to 14.10, the analysis of a precision experiment first computes, for each level separately, estimates of the repeatability variance s_r^2, the between-laboratory variance s_L^2 and the reproducibility variance s_R^2, as defined in 5.3, 5.4 and 5.5, and then the values of repeatability r and the reproducibility R.

7.3 The analysis, especially 7.1 a), includes a systematic application of statistical tests, a great variety of which are available from the literature and which could be used for the purposes of this International Standard. For practical reasons, only a limited number of these tests, as explained in section three, have been incorporated in this International Standard.

Section two : Organization of an inter-laboratory precision experiment

NOTE — 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 modified as appropriate 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. A member of the staff of that laboratory shall take full responsibility; he is called 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 available for the test method ?

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 10.5.2.)

i) How many laboratories should be recruited to cooperate in the experiment ? (See 10.1.)

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

k) 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 10.5.1.)

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

m) Who should be appointed to be executive officer ?

9.2 The tasks of the statistical expert are

a) to contribute his specialized knowledge in designing the experiment;

b) to analyse the data;

c) 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, in particular

a) to enlist the cooperation of the requisite number of laboratories and to ensure 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 draft instructions and circulate them to the supervisors early enough in advance for them to raise comments or queries;

d) to design suitable forms for the operator to use as a working record and for the supervisor to report the test results;

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

9.4 The tasks of the supervisor are

a) to hand out the samples to the operator(s) in keeping with the instructions of the executive officer;

b) to supervise the execution of the tests (the supervisor shall not take part in performing the tests);

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

9.5 The tasks of the operators are

a) to perform the tests in accordance with the standard test method;

b) to report any anomalies or difficulties experienced (see 10.5.1 and 10.5.3).

9.6 The final tasks of the panel are

a) to discuss the report by the statistical expert;

b) to establish final values for the repeatability and reproducibility;

c) to decide if further actions are required for improving the standard for the test method or with regard to laboratories whose results have been rejected as outliers [see 11.2.3 d)].

9.7 As 9.2 and 9.6 are considered during the final stages of the statistical analysis, they are discussed further 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. The use of at least six levels is desirable in order to establish the relationship between these quantities in a satisfactory manner. On the other hand, for the example on the determination of the softening point of a tar product given in clause 23 (with a range of levels from 88 to 102 °C), the use of four levels may be considered as more than adequate.

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 fewer than eight; and if only one level is of interest, the number of laboratories should preferably be higher, for example 15 or more.

Regarding the value of n, the recommended figure is two except where it is customary to make a larger number of replicates, such as 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 from all laboratories likely to use the test method. Volunteers may not represent a realistic cross-section of laboratories. However other practical considerations may intervene, for example, a requirement that participating laboratories be distributed over different continents or climatic regions may affect the pattern of representation. The panel should lay down the recruitment policy and the requirements for the participating laboratories.

10.2.2 In enlisting the cooperation of the requisite number of laboratories, their responsibilities should be clearly stated. An example of a suitable enlistment questionnaire that may be used for this purpose is given below :

Questionnaire on inter-laboratory study

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

.....

.....

1 Our laboratory is willing to participate in the precision experiment for this standard test method.

YES ☐ NO ☐ (tick the appropriate box)

2 As a participant, we understand that

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

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

c) the method shall be strictly adhered to;

d) samples shall be handled in accordance with instructions;

e) a qualified operator shall perform the test.

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

3 Comments : .....

.....

.....

Signature : .....

Company or laboratory : .....

10.3 Heterogeneity of the material

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

In general when publishing values of r and R, it is recommended that the material used in the precision experiment should be clearly specified along with the range of materials to which the values can be expected to apply.

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 shall be carried out under repeatability conditions, i.e. in a short interval of time and by the same operator, and without any intermediate recalibration of apparatus unless this is an integral part of making a determination.

c) It is not essential that the q groups of n tests each be performed strictly within a short interval; different groups of tests may be carried out on different days.

d) If in the course of the tests the operator should drop out, another operator may complete the tests, provided that the change does not occur within a group of n tests at one level but only occurs between two of the q groups. Any such change shall be reported with the test results.

e) It is essential that a group of n tests under repeatability conditions be performed independently as if they were n tests on different material. As a rule, however, the operator will know that he is testing identical material, but the point should be stressed to him in his instructions that the whole object of the experiment is to determine what differences in results can occur in actual testing. If it is feared that, despite this warning, previous results may influence succeeding test results and thus the repeatability variance, then a split-level experiment is considered the correct procedure (see 10.4.2).

10.4.2 An alternative procedure, sometimes adopted when n equals 2, is that of using split-level experiments. Adoption of this procedure may be considered when it is feared that the operator, when testing successive identical samples, may be influenced by the result of his first test. In this procedure, instead of using two samples that the operator has been told should be identical, or 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 small) and each of the p laboratories receives one sample from series **a** and one from series **b** for testing. It shall be distinguished clearly which test result belongs to series **a** and which to series **b**; they cannot be interchanged as can two test results on identical material. The values of r and R derived from a split-level experiment are valid for the mean level m equal to $(m_a + m_b)/2$.

The split-level design requires a slight modification in the statistical analysis, as discussed in section three.

10.4.3 Additional aspects of organizing the tests are as follows :

a) 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;

b) any preliminary checking of equipment should be as laid down in the standard method;

c) all samples should be clearly labelled with the name of the experiment and a sample identification.

10.5 Instructions to the operators

10.5.1 Before performing the tests the operators should receive no instructions that supplement those contained in the standard test method; these alone should suffice. The operators should, however, be encouraged to comment on the standard, in particular to state whether the instructions contained in it are sufficiently unambiguous and clear. For example, ambiguities may arise when a standard has been translated into different languages. However, it is desirable that all participating laboratories report their test results to the same number of decimal places, and the supervisors should be instructed accordingly. In commercial practice, the test results may be rounded rather crudely, and in a precision experiment it may be advisable to use one more decimal than is customary or laid down in the standard method. When r or R may depend on the level m, different rules for rounding may be needed for different levels.

10.5.2 An operator may not achieve normal precision when he carries out a test method for the first time or after a long interval. In such cases, subject to the decision of the panel or of the supervisors, the operators may be allowed to carry out a few unofficial tests in order to gain experience with the method before starting testing on the official samples of the precision experiment. Such preliminary familiarization tests shall never be performed on the official samples, and material for them should be supplied by the executive officer.

10.5.3 The operators should be told to report any occasions when they are not able to follow their instructions or when they accidentally fail 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 which should contain the following information :

a) the final test results, taking particular care to avoid transcription and typing errors, e.g. by using photocopies of the operators' results;

b) the original observations or readings (if any) from which the final results were derived, possibly by photocopying the operators' workbook;

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

d) information about irregularities or disturbances that may have occurred during the tests, including any change of operator that may have occurred along with a statement as to which tests were performed by which operator;

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

f) the date(s) on which each sample was tested;

g) information about the equipment used, if relevant;

h) any other relevant information.