GUIDE 33

Uses of certified reference materials

First edition 1989

Contents

International Organization for Standardization
Case postale 56 • CH-1211 Genève 20 • Switzerland

Printed in Switzerland

Foreword

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

ISO guides are intended essentially for internal use in ISO committees or in some cases for the guidance of member bodies when dealing with matters that would not normally be the subject of an International Standard.

ISO Guide 33 was drawn up by the ISO Committee on reference materials (REMCO) in collaboration with Technical Committee ISO/TC 69, *Applications of statistical methods*, and was submitted directly to ISO Council for acceptance.

Introduction : "Proper use" and "misuse" of CRMs

Today's world of modern technology requires a vast number of certified reference materials (CRMs) in widely diverse fields and the demand for such materials is expected to increase. The preparation of a CRM is a time-consuming, meticulous and expensive endeavour and consequently it has not always been, and will continue not to be, possible to satisfy the demand for all types and quantities of CRMs. For this reason, CRMs must be used properly, i.e. effectively, efficiently and economically.

Certified reference materials must be used consistently to ensure reliable measurements. However, in doing so, the magnitude of the supply of that RM, its relative cost, its availability (accessibility) and the measurement technique, be it destructive or non-destructive, must be considered. Also important to the user is the fact that the misuse of an RM may not provide the intended information.

Misuse of CRMs differs from incorrect use. The user of a CRM is expected to be familiar with all information pertinent to the use of the CRM as specified by the producer. He should comply with such factors as the period of validity of the CRM, the prescribed conditions for storage of the CRM, instructions for the use of the CRM, and specifications for validity of the certified properties of the CRM. A CRM should not be used for a purpose other than that for which it was intended. Nevertheless, from time to time, when a user must resort to applying a CRM in an incorrect manner because of the unavailability of a suitable CRM, he must be fully cognizant of the potential pitfalls and therefore assess his measurement output accordingly.

There are many measurement processes where CRMs are in general use but are replaceable by a host of working standards such as homogeneous materials, previously analysed materials, pure compounds, solutions of pure elements, etc. Some examples are where only a "rough" estimate of the trueness or precision of a method is sought, where "blind" unknown check samples are used routinely in quality control programmes, and where only the variation in trueness or precision of a method with some parameter such as time, analyst, instrument, etc., is being evaluated. The first example illustrates the use of a CRM where the well-defined certified value and uncertainty of the CRM is under-utilized. The others illustrate the case where a series of "one-time" trueness and precision assessments are compared with one another. There is no need to base that comparison on a well-defined certified value and uncertainty of a CRM. The advantages in using CRMs are that the user has the means to assess the trueness and precision of his measurement method and establishes metrological traceability for his results.

Whether the use of CRMs in these procedures is in fact "misuse" depends largely on the availability and relative cost of the CRMs. Where CRMs are in short supply or very expensive, their use would indeed be misuse. However, for CRMs in ample supply or where similar CRMs are available from one or more producers, it is strongly recommended that CRMs always be used instead of working standards because of the resultant enhanced confidence in the measurement output.

It is important that users remain aware that the preparation of working standards for use instead of CRMs has an associated cost based on factors such as material cost, facility usage charges, personnel labour rates, etc., in which the material cost is in general the lowest. For some CRMs such as the complex compositional materials certified for chemical composition, the cost of preparing working standards to match the composition of real samples can exceed that of available CRMs. In these cases, the use of CRMs is recommended.

The user should be aware of the potential misuse of CRMs as "blind" unknown check samples in quality control programmes. Where there are only a few CRMs in an area of expertise, they are easily recognized and they may therefore not satisfy the intended purpose. Moreover, the same CRMs should never be used for both calibration purposes and as "blind" unknown check samples in a measurement process.

The misuse of CRMs can occur also when the user does not fully take into account the uncertainty in the certified property. The overall uncertainty of a certified property of a CRM can have contributions from the inhomogeneity of the material, the within-laboratory uncertainty and, where applicable, the between-laboratories uncertainty. The level of homogeneity defined for a CRM by the producer is dependent on both the statistical design used to evaluate it and the repeatability of the method of measurement. For certain CRMs, the level of homogeneity is valid for a test portion defined by mass, physical dimension, time of measurement, etc. The user must be aware that the use of a test portion that does not meet or exceed that specification could severely increase the contribution of the inhomogeneity of the CRM to the uncertainty of the certified property to the point where the statistical parameters of certification are no longer valid.

The variation in the repeatability of different methods has another implication for the user. Since the degree of inhomogeneity of a CRM is dependent on the repeatability of the method of measurement, it is possible that a user, in applying a method capable of better repeatability, could detect inhomogeneity in that CRM. In such cases, the observed inhomogeneity is already accounted for in the statistical parameters for the certified property and therefore the statistical tests presented in this Guide remain valid but the scientific basis for using that particular CRM to give a true assessment of the user's method must again be questioned.

It is well known that the methods of measurement of a property are not capable of equal repeatability. Accordingly there could arise instances where the user may wish to assess a method that has greater repeatability than that or those used in the certification of the CRM. In such cases, the statistical tests presented in this Guide remain valid but the scientific basis for using that particular CRM to give a true assessment of the precision (and possibly the trueness) normally expected from the user's method must be questioned. It is recommended that the user resort to a CRM of lesser uncertainty, if available.

For RMs certified by a "definitive" method, the user should not assume that his method is capable of matching the precision and trueness reported for the CRM. It is unreasonable therefore to apply the statistical procedures in this Guide for assessing the trueness and precision of a method by application to a CRM using the certification parameters for a property reported by the producer. The user, as a consequence, must either experimentally establish or make estimates based on available information for those parameters that are more appropriate. Similarly, where a user applies a method to an RM that has been certified by a single different method, the user must not assume that the certification parameters for the certified property are applicable to his method except in cases where the trueness and precision capable by both methods are known to be comparable.

One of the important considerations in selecting a CRM for use either in assessing the trueness and precision of a method or in the calibration of instruments in a method is the level of uncertainty required by the end-use of the method. Obviously the user should not apply a CRM of greater uncertainty than permitted by the end-use. This Guide states that the uncertainty of the replicated measurement of a CRM is twice the between-laboratories standard deviation of the certification programme, when this parameter is known, or four times the repeatability of the method on the CRM. These values represent the lowest level of uncertainty achievable with this CRM and may be used as a guide in determining whether it satisfies the uncertainty requirement of the end-use.

The selection of CRMs must take into account not only the level of uncertainty required for the intended purpose but also their availability, cost, and chemical and physical suitability for the intended purpose. For example, the unavailability or high cost of one CRM could force a user to resort to using another CRM of greater uncertainty than the preferred one. Also, in chemical analysis, a CRM of greater, but still acceptable, uncertainty in the certified property may be preferred over another CRM because of better matching with the composition of real samples. This could result in minimizing "matrix" or chemical effects in the measurement process which are capable of causing errors far greater than the difference between the uncertainties of the CRMs.

In conclusion, CRMs are meant to fulfil many purposes. Accordingly, a CRM used properly for one purpose in one laboratory may be misused for another purpose in another laboratory. It is recommended to the user that he consider the suitability of a CRM for his intended purpose on a case-by-case basis.

Uses of certified reference materials

Section 1 : General

1.1 Scope

Section 1 of this Guide presents definitions (with indication of their sources) of terms used, and sets out the statistical considerations on which the Guide is based.

Section 2 presents recommendations for developing criteria for the assessment of the precision and trueness of a measurement process by the use of CRMs. It pertains only to CRMs characterized to be homogeneous as described in ISO Guide 35[1]. The use of CRMs is essential for assessment of trueness and optional for assessment of precision.

Section 3 discusses the use of CRMs for the definition and realization of conventional measurement scales.

1.2 Definitions

1.2.1 measurement process : All the information, equipment and operations relevant to a given measurement.

NOTE — This concept embraces all aspects relating to the performance and quality of the measurement; it includes, for example, the principle, method, procedure, values of the influence quantities and the measurement standards.

[VIM : 1984[2]]

1.2.2 influence quantity : A quantity which is not the subject of the measurement but which influences the value of the measurand or the indication of the measuring instrument.

EXAMPLES — ambient temperature; frequency of an alternating measured voltage.

[VIM : 1984[2]]

1.2.3 reference material (RM) : A material or substance one or more properties of which are sufficiently well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials.

[ISO Guide 30 : 1981[3]]

1.2.4 certified reference material (CRM) : A reference material one or more of whose property values are certified by a technically valid procedure, accompanied by or traceable to a certificate or other documentation which is issued by a certifying body.

[ISO Guide 30 : 1981[3]]

1.2.5 precision : The closeness of agreement between independent test results obtained under prescribed conditions.

[ISO 3534-1[4]]

1.2.6 repeatability : Precision under repeatability conditions.

[ISO 3534-1[4]]

1.2.7 repeatability conditions : Conditions where 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.

[ISO 3534-1[4]]

1.2.8 repeatability standard deviation : The standard deviation of test results obtained under repeatability conditions.

NOTE — It is a measure of the dispersion of the distribution of test results under repeatability conditions.

[ISO 3534-1[4]]

1.2.9 repeatability limit, r : The value less than or equal to which the absolute difference between two single test results obtained under repeatability conditions is expected to be with a probability of 95 %.

[ISO 3534-1[4]]

1.2.10 reproducibility : Precision under reproducibility conditions.

[ISO 3534-1[4]]

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

[ISO 3534-1[4]]

1.2.12 reproducibility limit, R : The value less than or equal to which the absolute difference between two single test results obtained under reproducibility conditions is expected to be with a probability of 95 %.

[ISO 3534-1 [4]]

1.2.13 bias : The difference between the expectation of the test results and an accepted reference value.

NOTE — Bias is a systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value.

[ISO 3534-1 [4]]

1.2.14 trueness : The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value.

NOTE — The measure of trueness is usually expressed in terms of bias.

[ISO 3534-1 [4]]

1.2.15 estimation, estimate, estimator (of parameters)

1.2.15.1 estimation : The operation of assigning, from the test result in a sample, numerical values to the parameters of a distribution chosen as the statistical model of the population from which this sample is taken.

[ISO 3534-1 [4]]

1.2.15.2 estimate : The result of an estimation.

[ISO 3534-1 [4]]

1.2.15.3 estimator : A statistic used to estimate a population parameter.

[ISO 3534-1 [4]]

1.3 Statistical considerations

1.3.1 Basic assumptions

All statistical methods used in this Guide are based on the following assumptions.

a) The certified value is the best estimate of the true value of the property of the CRM.

b) All variation, be it associated with the material (i.e. homogeneity) or the measurement process, is random and follows a normal probability distribution. The values of probabilities stated in this Guide assume normality. They may be different if there is deviation from normality.

1.3.2 Decision errors

The assessment of a measurement process on the basis of precision and trueness is always subject to rendering an incorrect conclusion because of

a) the uncertainty of measurement results and

b) the limited number of replicate results usually performed.

Increasing the number of measurements decreases the chance of an incorrect conclusion but, in many instances, the risk of making a wrong conclusion has to be balanced in economic terms against the cost of increasing the number of measurements. Accordingly, the rigour of the criteria developed for assessing a measurement process must take into account the level of precision and trueness requisite for the end-use.

For the purposes of this Guide, it is necessary to define the term "null hypothesis".

> The null hypothesis is the hypothesis to be accepted or rejected based on the outcome of the measurement. In this case the null hypothesis is that the measurement process has bias no greater than the limit chosen by the experimenter and variances no greater than the predetermined value; the alternative hypothesis is the hypothesis which is opposed to the null hypothesis (see also ISO 3534-1 [4]).

There are two types of possible error in accepting or rejecting the null hypothesis :

a) **error type I** : The error committed in rejecting the null hypothesis when in reality the null hypothesis is true.

   type I risk : The probability of committing error type I. Its value varies according to the real situation. The maximum value is called the significance level.

   significance level : The given value, usually designated by α, which limits the probability of committing error type I.

b) **error type II** : The error committed in failing to reject the null hypothesis when in reality the alternative hypothesis is true.

   type II risk : The probability, usually designated by β, of committing error type II. Its value depends on the real situation and can be calculated only if the alternative hypothesis is adequately specified.

   power of test: The probability of not committing error type II, usually designated by $(1 - \beta)$. It is the probability of rejecting the null hypothesis when in reality the alternative hypothesis is true.

The choice of the values of both α and β is usually based on economic considerations dictated by the importance of the consequences of the decision. These values as well as the alternative hypothesis should be chosen before the start of the measurement process.

Section 2 : Assessment of a measurement process

2.1 The cases to be considered

2.1.1 One laboratory

This is a check of precision and/or trueness of a measurement method as applied by one particular laboratory. The laboratory uses a CRM to check its measurement process for any particular reason at any time.

2.1.2 Interlaboratory programme

In this case the test procedure is performed by a number of laboratories as part of an organized programme, for example as described in ISO 5725[5]. The purpose of this programme is to establish the performance characteristics of a measurement process, against which a typical laboratory can compare its performance.

2.2 Requirements of limits

In order to satisfy the requirement, the measurement process must produce results with precision measure and/or trueness within the predetermined limits when it is applied to a CRM. The limit of precision is usually expressed in terms of standard deviation and the trueness requirement is expressed in terms of the bias of the measurement results against the certified value. These limits may originate from various sources.

2.2.1 Legal limits

Legal limits are those limits which are required by statute or regulation; for example, procedures for the analysis of sulfur dioxide in air are required to have a certain precision and trueness.

2.2.2 Accreditation schemes

In most cases the limits of bias and precision are consensus values agreed upon between the various participants concerned, e.g. producer, consumer and independent. For this reason, in most cases, these limits are derived from some realistic values, e.g. those obtained from the results of the certification campaign of the CRM, international tests of ISO standards, etc.

2.2.3 User of the process

This is where the laboratory, or the organization of which the laboratory is a part, imposes upon itself the limits of bias and precision, e.g. limits imposed by commercial requirements.

2.2.4 Previous experience

This is where the limits of bias and precision of the measurement process to be tested should be based on the values obtained from previously established measurement processes.

2.3 Choices of CRM

2.3.1 Relevance to measurement process

The user of the CRM must decide what properties of the CRM are relevant to his measurement process, taking into account the statement on intended use and instructions for the correct use of the CRM on the certificate.

a) **Level.** The CRM should have properties at the level appropriate to the level at which the measurement process is intended to be used, e.g. concentration.

b) **Matrix.** The CRM should have a matrix as close as possible to the matrix of the material to be subjected to the measurement process, e.g. carbon in low-alloy steel or carbon in stainless steel.

c) **Form.** The CRM may be a solid, liquid or gas. It may be a test piece or a manufactured article or a powder. It may need preparation.

d) **Quantity.** The quantity of the CRM must be sufficient for the entire experimental programme, including some reserve if it is considered necessary. Avoid having to obtain additional CRM later.

e) **Stability.** Wherever possible the CRM should have stable properties throughout the experiment. There are three cases :

1) the properties are stable and no precautions are necessary;

2) where the certified value may be influenced by storage conditions, the container should be stored, both before and after its opening, in the way described on the certificate;

3) supplied with the CRM is a certificate to define the properties (which are changing at a known rate) at specific times.

f) **Acceptable uncertainty of the certified value.** The uncertainty of the certified value should be compatible with the precision and trueness requirements outlined in 2.2.

2.3.2 Type of certification of CRM

The choice of the type of certification of the CRM is governed by the information required for the experimental programme. Refer to ISO Guide 35[1].

2.4 Carrying out the experiment

The procedure for the measurement must be fixed, i.e. a written document must exist laying down all the details. There must be no changes to the procedure during the course of the experiment.

2.4.1 Check of precision and trueness of a measurement process by one laboratory

Checking of precision of a measurement process as applied by a laboratory involves comparison between the within-laboratory standard deviation under repeatability conditions (or other defined conditions) and the required value of standard deviation.

Checking of trueness of a measurement process as applied by a laboratory involves comparison between the mean of the measurement results and the certified value of the CRM. The between-laboratories component of precision of the measurement process should be taken into account when making this comparison.

2.4.1.1 Number of replicate measurements, n

The number of replicate measurements required mainly depends on the values of α and β and the alternative hypothesis chosen for the assessment of precision.

Table 1 shows the relation between the degrees of freedom ν (where in this case $\nu = n - 1$) and the ratio of the within-laboratory standard deviation of the measurement process, σ_{w}, and the required value of the within-laboratory standard deviation, σ_{wo}, for various values of β at $\alpha = 0{,}05$. For example, for $n = 10$ the probability that the variance of the measurement results will pass the appropriate χ^2-test at $\alpha = 0{,}05$ is no more than 1 % when the within-laboratory standard deviation, σ_{w}, of the measurement process is equal to or larger than 2,85 times the required value of σ_{wo}.

Table 1 — Ratio of the standard deviation of the measurement process to the required value for various values of β and degrees of freedom ν at $\alpha = 0{,}05$

ν	$\alpha = 0{,}05$			
	$\beta = 0{,}01$	$\beta = 0{,}05$	$\beta = 0{,}1$	$\beta = 0{,}5$
1	159,5	31,3	15,6	2,73
2	17,3	7,64	5,33	2,08
3	6,25	4,71	3,66	1,82
4	5,65	3,65	2,99	1,68
5	4,47	3,11	2,62	1,59
6	3,80	2,77	2,39	1,53
7	3,37	2,55	2,23	1,49
8	3,07	2,38	2,11	1,45
9	2,85	2,26	2,01	1,42
10	2,67	2,15	1,94	1,40
12	2,43	2,01	1,83	1,36
15	2,19	1,85	1,71	1,32
20	1,95	1,70	1,59	1,27
24	1,83	1,62	1,52	1,25
30	1,71	1,54	1,46	1,22
40	1,59	1,45	1,38	1,19
60	1,45	1,35	1,30	1,15
120	1,30	1,24	1,21	1,11

2.4.1.2 The CRM

The user should confirm the suitability of the CRM with respect to certified value with its uncertainty, method of characterization, date of certification, statement of intended use, expiration date (particularly for a relatively unstable CRM), packaging and storage conditions and special instructions for correct use given in the certificate and the size of test portion required for the measurement process.

2.4.1.3 Measurement

The user should perform independent replicate measurements. "Independent", in a practical sense, means that a replicate result is not influenced by previous replicate results. To perform replicate measurements means to repeat the whole procedure. For example, in the chemical analyses of a solid material, the procedure should be repeated from the weighing of the test portion to the final reading or calculating of the result. Taking aliquots from the same sample solution is not independent replication.

Independent replicate measurements can be achieved in various ways depending on the nature of the process. In some, however, parallel replication is not recommended because an error committed at any step of the procedure could affect all replicates. For example, in the case of iron ore analyses, replication of the analytical procedure is carried out at different times and includes appropriate calibration.

The measurement results could, if necessary, be scrutinized for possible outliers using the rules described in ISO 5725[5]. It should be noted that an excessive number of suspected outliers indicates problems in the measurement process.

2.4.1.4 Assessment of precision

The precision of the measurement process is assessed by comparing the within-laboratory standard deviation under repeatability conditions with the required value of the within-laboratory standard deviation, σ_{wo}.

Compute the average, $\bar{x}$, and standard deviation, s_{w} :

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \ldots (1)$$

$$s_{\mathrm{w}} = \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n - 1} \right]^{1/2} \qquad \ldots (2)$$

Compute the following ratio :

$$\chi_{\mathrm{c}}^2 = \left(\frac{s_{\mathrm{w}}}{\sigma_{\mathrm{wo}}} \right)^2 \qquad \ldots (3)$$

where

x_i is the individual result;

n is the number of results excluding outliers;

σ_{wo} is the required value of the within-laboratory standard deviation;

$$\chi^2\text{-table} = \frac{\chi^2_{(n-1);\,0,95}}{n-1}$$

= 0,95-quantile of the χ^2-distribution at degrees of freedom $(n - 1)$ divided by the degrees of freedom $(n - 1)$

Decision :

$\chi^2_c \leqslant \chi^2$-table : There is no evidence that the measurement process is not as precise as required.

$\chi^2_c > \chi^2$-table : There is evidence that the measurement process is not as precise as required.

2.4.1.5 Assessment of trueness

The trueness of the measurement process is checked by comparing the average $\bar{x}$ with the certified value, μ.

There are two factors contributing to the difference between the certified value and the measurement results :

1) the error of the certified value;

2) the error of the results of the measurement process being assessed expressed by its standard deviation σ_D.

For a CRM prepared in accordance with ISO Guide 35[1], the uncertainty of the certified value should be small in comparison with σ_D. The following general formula is used as the criterion for acceptance :

$$-a_2 - 2\sigma_D \leqslant \bar{x} - \mu \leqslant a_1 + 2\sigma_D \quad \ldots (4)$$

where a_1 and a_2 are adjustment values chosen by the experimenter according to economic or technical limitation or stipulation.

The standard deviation associated with the measurement process, σ_D, arises from the fact that a measurement procedure performed on the same material does not, in general, yield identical results every time it is applied. This fluctuation is attributed to unavoidable random errors inherent in every measurement process because the factors that may influence the outcome of a measurement cannot all be completely controlled. This random fluctuation of the measurement results should be taken into account when assessing the trueness of the procedure. For this purpose, the random fluctuation can be divided into two parts :

a) within-laboratory, or short-term fluctuation, which has a mean of zero and standard deviation of σ_w; an estimate of σ_w is given as s_w in equation (2);

b) between-laboratories fluctuation, which has a mean of zero and standard deviation of σ_{Lm}. This fluctuation is caused by one or a combination of various factors such as operators, equipment, laboratories, time, etc. When the assessment experiment is performed by only one laboratory, σ_{Lm} cannot be determined directly. In many cases it is sufficient to substitute σ_{Lm} by the long-term within-laboratory standard deviation. Otherwise, σ_L supplied by the certificate of the CRM, or from other sources such as an appropriate International Standard, can be used to replace σ_{Lm}.

The value of σ^2_D is therefore given as the sum of uncertainties :

$$\sigma^2_D = \sigma^2_{Lm} + \frac{s^2_w}{n} \quad \ldots (5)$$

where n is the number of replicate determinations performed for the assessment of the measurement process by the assessing laboratory.

For many measurement processes, σ_w is small in comparison with σ_{Lm}; consequently for large numbers of replications $(n > 10)$, σ_D in equation (5) can be equated with σ_{Lm} or σ_L. Thus, in this case equation (4) can be simplified :

$$-a_2 - 2\sigma_{Lm} \leqslant \bar{x} - \mu \leqslant a_2 + 2\sigma_{Lm} \quad \ldots (6)$$

2.4.1.6 Example : Analysis for the iron content in iron ores

Purpose of investigation :

To check whether a certain analytical method (method A) is sufficiently precise and accurate by using an iron ore CRM for the case where $a_2 = a_1 = 0$.

Certificate information :

The available CRM was certified by an interlaboratory programme for 13 elements, including iron.

μ = 60,73 % Fe

σ_{wo} = 0,09 % Fe

σ_L = 0,20 % Fe

Analysis : $n = 11$

x_i(% Fe) = 60,7 60,8 60,8 60,9 60,9
60,9 61,0 61,0 61,1 61,2 61,9

(the x_i values have been arranged in ascending order).

Dixon test for outlier (ISO 5725) :

$x(11)$ is a suspect.

$$Q = \frac{x(11) - x(10)}{x(11) - x(2)} = \frac{61,9 - 61,2}{61,9 - 60,8} = \frac{0,7}{1,1} = 0,636$$

The critical value for $n = 11$ at 5 % is 0,502 and at 1 % is 0,60. Therefore, $x(11)$ is an outlier and should be rejected. The remaining data are to be used for further computation. The new n is 10.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = 60{,}930 \text{ \% Fe}$$

$$s_w = \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \right]^{1/2} = 0{,}149 \text{ \% Fe}$$

$$\chi^2 = \left(\frac{s_{\mathrm{w}}}{\sigma_{\mathrm{wo}}}\right)^2 = \left(\frac{0{,}149}{0{,}090}\right)^2 = 2{,}76$$

$$\chi^2\text{-table} = \frac{\chi^2_{9;\,0{,}95}}{9} = 1{,}88$$

$$\chi^2_{\mathrm{c}} > \chi^2\text{-table}$$

As a conclusion, the within-laboratory standard deviation of method A is not as good as required. The method should be investigated chemically.

Second assessment :

The second set of analytical results, after method improvement, is :

$n = 10$

x_i = 60,94 60,99 61,04 61,06 61,06
61,09 61,10 61,14 61,21 61,24

Visual observation of the results shows no reason to suspect that there is an outlier; therefore the Dixon test is not necessary.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = 61{,}087\ \%\ \mathrm{Fe}$$

$$s_{\mathrm{w}} = \left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}\right]^{1/2} = 0{,}092\ \%\ \mathrm{Fe}$$

$$\chi^2_{\mathrm{c}} = \left(\frac{s_{\mathrm{w}}}{\sigma_{\mathrm{wo}}}\right)^2 = \left(\frac{0{,}092}{0{,}090}\right)^2 = 1{,}04 < \chi^2\text{-table}$$

$$|\bar{x} - \mu| = 61{,}087 - 60{,}730$$

$$= 0{,}357\ \%\ \mathrm{Fe}$$

$$2\sigma_{\mathrm{L}} = 0{,}40\ \%\ \mathrm{Fe}$$

$$|\bar{x} - \mu| < 2\sigma_{\mathrm{L}}$$

Therefore the method is as accurate as required.

2.4.2 Assessment of a measurement process by an interlaboratory measurement programme

One of the most important criteria that a measurement process must satisfy in order to receive "widely accepted" or "standard" status is that it is capable of producing results with precision and trueness sufficient for the end-use when applied by a qualified operator. In most instances, the precision and trueness of such a candidate process are assessed by an interlaboratory measurement programme in which the participants are selected so as to provide a representative sample of the laboratories which will ultimately apply that measurement process. The procedure of conducting an interlaboratory measurement programme is described in ISO 5725[5].

2.4.2.1 Number of laboratories, k, and number of replicate measurements per laboratory, n

Ideally the values of k and n should be selected according to the limit of bias between the certified value of the CRM and the value obtained by the interlaboratory measurement programme, M, the significance level, α, and the type II risk, β. In many cases, the choice of k and n is limited by the availability of participating laboratories. The detailed procedure for computing the ideal values of k and n is described in 2.4.2.6.

2.4.2.2 Experimental

An interlaboratory measurement programme is often conducted as part of an experiment to estimate precision. A detailed procedure for performing such an experiment is described in ISO 5725[5].

2.4.2.2.1 Check and distribution of the CRM

a) The CRM should be checked as described in 2.4.1.2.

b) Where subdivision of the unit of the CRM occurs prior to distribution, it must be performed with care to avoid any additional error. Relevant International Standards on sample division should be consulted. If the unit of the CRM has a fixed form, e.g. metal disc, the units should be selected on a random basis for distribution. If the measurement process is non-destructive, it is possible that all laboratories in the interlaboratory measurement programme be given the same unit of the CRM but this will extend the time-frame of the programme.

2.4.2.2.2 Measurement

The coordinator of the interlaboratory measurement programme must specify n, the number of independent replicate determinations to be performed by each laboratory, and the organizational factors of interlaboratory programmes such as time limit for submission of results, the size of test portion, etc.

Methods for computing the precision measures from the results of an interlaboratory programme are described in ISO 5725[5].

2.4.2.3 Assessment of precision

The precision of the measurement process as applied to the CRM is expressed in terms of s_{w}, the estimate for the within-laboratory standard deviation, and s_{Lm}, the estimate for the between-laboratories standard deviation.

2.4.2.3.1 Within-laboratory precision

The estimates of the within-laboratory standard deviation of the interlaboratory comparison, σ_{w}, can be compared with the required value of σ_{wo} in a manner analogous to that described in 2.4.1.4 by :

$$\chi^2_{\mathrm{c}} = \left(\frac{s_{\mathrm{w}}}{\sigma_{\mathrm{wo}}}\right)^2 \qquad \ldots (3)$$

χ^2_{c} is compared with $\chi^2\text{-table} = \dfrac{\chi^2_{k(n-1);\,0{,}95}}{k(n-1)}$

Decision :

$\chi_c^2 \leqslant \chi^2$-table : There is no evidence that the within-laboratory precision of the measurement process is not as good as required.

$\chi_c^2 > \chi^2$-table : There is evidence that the within-laboratory precision of the measurement process is not as good as required.

2.4.2.3.2 Between-laboratories precision

The between-laboratories precision can be assessed indirectly by testing the following statistic :

$$\chi_c^2 = \frac{s_w^2 + n s_{Lm}^2}{\sigma_{wo}^2 + n \sigma_L^2} \qquad \ldots (7)$$

For many test methods, the within-laboratory standard deviation is equal to or smaller than the between-laboratories standard deviation; therefore equation (7) can be simplified :

$$\chi_c^2 = \frac{n s_{Lm}^2}{n \sigma_L^2} \qquad \ldots (7a)$$

χ_c^2 is compared with χ^2-table $= \dfrac{\chi^2_{(k-1);0,95}}{k-1}$

Decision :

$\chi_c^2 \leqslant \chi^2$-table : There is no evidence that the between-laboratories standard deviation of the measurement process is not as good as required.

$\chi_c^2 > \chi^2$-table : There is evidence that the between-laboratories standard deviation of the measurement process is not as good as required.

2.4.2.4 Assessment of trueness

The trueness of the measurement process is checked by comparing the overall mean of the interlaboratory measurement programme, $\bar{\bar{x}}$, with the certified value of the CRM. In a manner analogous to 2.4.1, the criterion for acceptance is :

$$-a_2 - 2\sigma_D \leqslant \bar{\bar{x}} - \mu \leqslant a_1 + 2\sigma_D \qquad \ldots (8)$$

where σ_D is the standard deviation of the overall mean of the interlaboratory comparison for the measurement process, and is given by :

$$\sigma_D^2 = \frac{s_{Lm}^2 + s_w^2/n}{k} \qquad \ldots (9)$$

Decision :

1) If equation (8) is satisfied, there is no evidence that the bias of the measurement process exceeds the prescribed limit including the adjustment value.

2) If equation (8) is not satisfied, there is evidence that the bias of the measurement process exceeds the prescribed limit including the adjustment value.

2.4.2.5 Example

A CRM was used for assessing the precision and trueness of an analytical method by interlaboratory comparison.

μ = 60,73 % Fe
σ_w = 0,09 % Fe
σ_L = 0,020 % Fe

Method assessed :

Iron ores — Determination of total iron content — Pollution-free method (Perchloric acid oxidation)

Interlaboratory programme :

Number of participating laboratories : 35
Number of results : 113 (some laboratories did not report the specified number of replicate determinations)

Let us suppose that the programme coordinators decide that a limit of bias of ± 0,08 % of Fe is technically reasonable and/or acceptable; then $a_1 = a_2 = 0{,}08$.

Result of evaluation :

Outliers

One set of laboratory results (two results) was identified as an outlier due to poor precision and was excluded from further computation.

k = 34
N = 111
$n = N/k = 3{,}26$
$\bar{\bar{x}}$ = 60,67 % Fe
s_w = 0,10 % Fe $\qquad k(n-1) = 77$
s_{Lm} = 0,06 % Fe $\qquad (k-1) = 33$

Statistical tests :

Precision

a) Within-laboratory

$$\chi_c^2 = \left(\frac{s_w}{\sigma_{wo}}\right)^2 = \left(\frac{0{,}010}{0{,}09}\right)^2 = 1{,}23$$

$$\chi^2\text{-table} = \chi^2_{77;0,95} = 1{,}28$$

$$\chi_c^2 < \chi^2\text{-table}$$

b) Between-laboratories

$$\chi_c^2 = \frac{0{,}10^2 + 3{,}26 \times 0{,}06^2}{0{,}09^2 + 3{,}36 \times 0{,}20^2}$$

$$= \frac{0{,}021\,7}{0{,}142\,5} = 0{,}152\,5 < 1$$

The method is as precise as those used in the certification of the CRM.

c) Bias

$$|\bar{\bar{x}} - \mu| = |60{,}73 - 60{,}67| = 0{,}06 \text{ \% Fe}$$

$$\sigma_D^2 = \frac{s_{LM}^2 + s_w^2/n}{k}$$

$$= \frac{0{,}06^2 + (0{,}10^2/3{,}26)}{34} = 0{,}000\,2$$

$$\sigma_D = 0{,}014$$

$$a_1 + 2\sigma_D = 0{,}108$$

$$|\bar{\bar{x}} - \mu| < (a_1 + 2\sigma_D)$$

There is no evidence that the bias exceeds the prescribed limit including the adjustment value.

2.4.2.6 Detailed procedure for computing the ideal combination of k and n

The overall mean of the results of the interlaboratory comparison programme represents the best estimate of the value of the property of the CRM as determined by the measurement process being investigated. In practice the trueness of the measurement process is defined as the agreement between this overall mean value, $\bar{\bar{x}}$, and the certificate value of the CRM, μ. The estimate of the bias of the measurement process, δ, is given by :

$$\delta = \bar{\bar{x}} - \mu \qquad \ldots (10)$$

and its variance :

$$\sigma_\delta^2 = V[\bar{\bar{x}}] = \frac{\sigma_{Lm}^2 + \sigma_w^2}{k} \qquad \ldots (11)$$

The null hypothesis is that the measurement process is biased by a_1 in the positive direction or a_2 in the negative direction; $E(\delta) = a_1$ in the case of positive bias or $E(\delta) = -a_2$ in the case of negative bias. Figure 1 shows that, if the null hypothesis is correct, there is up to α chance (100α %) that the measurement process will be rejected as being unacceptably biased. For a value of $\alpha = 0{,}05$, the acceptance criterion is given in equation (8).

For simplification, only the positive bias will be discussed; the negative bias follows the same arguments. The alternative hypothesis is that the measurement process is in fact biased by M_1, $E(\delta) = M_1$; $E(\delta)$ is the expected value of δ. If this is the case, the dotted line in figure 1 shows that the chance that the measurement process is accepted to have allowable bias is (100β %). For $\beta = 0{,}05$ (5 %), and $\alpha = 0{,}05$ (5 %), the corresponding value of M_1 will be

$$M_1 = 4\sigma_\delta + a_1 \qquad \ldots (12)$$

and

$$M'_1 = M_1 - a_1 = 4\sigma_\delta \qquad \ldots (13)$$

where M'_1 is the minimum value of excessive bias in the measurement process that can be detected on the basis of the interlaboratory results with $\alpha = 0{,}05$ and $\beta = 0{,}05$.

For a given property of the CRM, the value of the particular M'_1 decreases as $V[\bar{\bar{x}}]$ decreases (as n and k increase). Thus, for a desired value of M'_1, a set of combinations of k and n can be computed by equations (10) to (13). One of these combinations should be chosen for the experiment. This chosen combination should also be checked to determine if it satisfies the requirement for a precision experiment (ISO 5725[5]).

2.5 General remarks

The criteria used for assessment of a measurement process described in this Guide are the limits for both precision and bias. In order that these criteria are workable, they must be compatible both with the CRM and the state of the art of the measurement process. The following factors should be considered when determining these limits.

2.5.1 The CRM

The certified value of a property of a CRM is the best estimate of the true value that the certifying body can obtain. This estimate has certain degrees of uncertainty. For a CRM this uncertainty is in the form of bias, i.e. the difference between the certified value and the true value and random variation due to inhomogeneity of the material. For a good CRM, the inhomogeneity should be generally negligible. In many industries, one often uses a group of CRMs of the same type. In this case the bias of an individual CRM becomes random uncertainty, i.e. there exists between-CRMs variation. The magnitude of this variation cannot be determined exactly. It can only be estimated from the knowledge of the measurement process. Nevertheless, it must be taken into consideration when determining the required limits for precision and bias.

2.5.2 The measurement process

For the purposes of this Guide, the precision of a measurement process can be subdivided into three parts : the short-term within-laboratory precision, long-term within-laboratory precision, and between-laboratories precision.

The required limits for both precision and bias must be determined in accordance with this partitioning. In some cases, part of the measurement uncertainty changes from random into systematic. For example, for a particular laboratory, part of the between-laboratories variance of the process becomes part of the systematic error of the laboratory. This part of the uncertainty that is considered as random on an interlaboratory experiment becomes systematic in a particular laboratory calibration. This factor must be taken into account when determining the limits.

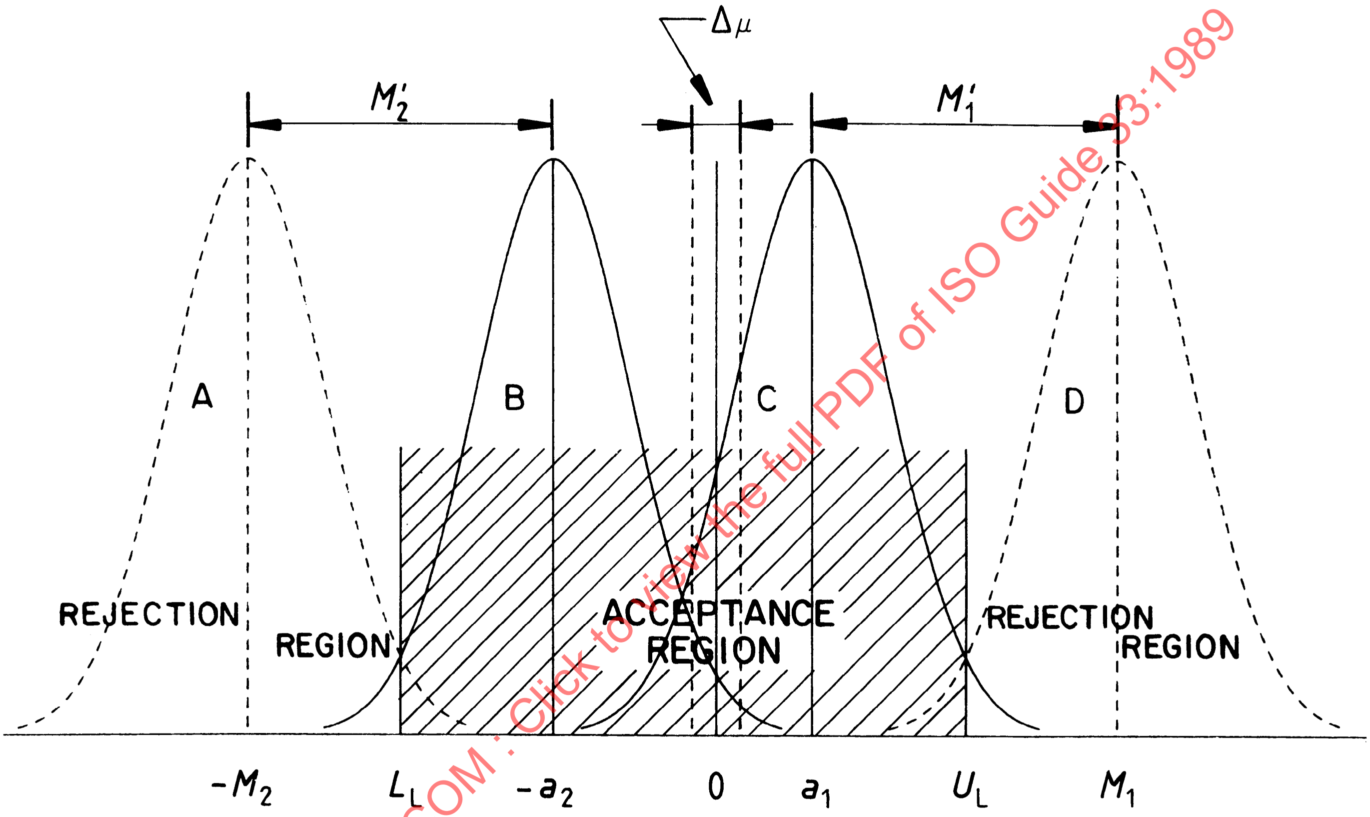

Figure 1 — Frequency distribution of $\delta = \bar{\bar{x}} - \mu$

B and C are lower and upper limits of distribution of acceptable δ. A and D are lower and upper alternative distribution of δ.

Section 3 : Defining and realizing conventional scales

3.1 General principles

Many measurement scales have been used since the earliest civilizations. Originally almost all of them were conventional, independent and inaccurate. Scientific and technical progress as well as international trade have brought both the need and the possibility of a unique, rational, self-consistent international system of units, SI, which has been officially adopted worldwide. Nevertheless, it is not applicable to certain types of measurements for which it is necessary to create, sustain and use certain conventional units which are not within the scope of SI. In other cases the unit relating to the quantity to be measured lies within the frame of SI, but the reproduction of the unit according to the definition is technically difficult and expensive. The realization of the measurement is therefore more convenient on a practical scale of reference values assigned to material properties.

A well-known example of this type of scale is the International Practical Temperature Scale. The fixed points of this scale are linked to definitive measurements carried out in metrological laboratories. Though a reference value scale and a pure conventional scale differ theoretically from each other, they are similar with respect to the use of reference materials, and they will therefore be discussed together as conventional scales.

Conventional scales are based on the values assigned to reference materials. The assigned values are stated in standard specifications, international recommendations or other reference documents; therefore a reference material realizing a fixed point on a conventional scale should have the same quality all over the world. RMs of this type are certified for property values, i.e. they are measured on standard equipment with reference methods at metrological or other authorized laboratories.

It is evident that the RMs ensure only the fixed points of a measurement scale. Measurement on a scale requires either a fixed point and a mathematical function passing through it, or two or more fixed points with stated means of interpolation between them.

NOTE — There exist some special discontinuous scales, e.g. the Moh's scale for measuring hardness in geological tests. The scale is based on ten minerals to which are assigned ten grades of hardness : each harder mineral scratches the less hard one.

A conventional scale has two fundamental pillars : the reference material, realizing the fixed point(s), and the standard specification (or similar document), giving the method of measurement. Both of them should be strictly defined to ensure the compatibility of measurements on the conventional scale.

The standard specification provides detailed information necessary to establish and use a scale based on assigned values or it may provide protocols for the experimental and calculational procedures to be used in measurements which depend on assumptions. It is advisable to prescribe the requirements of the reference material in the same standard specification as that in which the method of measurement is described. By means of the necessary RMs and relevant standard specifications, the user can realize the measurement scale and with the aid of such a scale can measure his sample or calibrate his instrument.

To estimate the uncertainty of a measurement on the scale, the user should consider the uncertainties in the creation of the scale and the uncertainty associated with the realization of its fixed points by the RM. Sometimes the users demand a level of uncertainty in the end-use which is lower than the uncertainty of the fixed points defined by the CRM (e.g. in measurement of the pH of blood). They must realize that the uncertainty of the measurements on the scale is necessarily greater than the uncertainty of the fixed points. According to section 3 of this Guide, even the uncertainty of the replicated measurement of a CRM is four times the repeatability of the method on the CRM, and the setting-up of a scale (the appropriate selection of the points, the characteristics and repeatability of the interpolating instrument, etc.) also contributes to the overall uncertainty.

The selection of CRMs for realizing the fixed points of a scale should be directed by the required level of uncertainty of the end use. To minimize the uncertainty of the measured value on the scale, the user should employ CRMs which have been certified in terms of the units of the scale. Obviously, the user is expected to be familiar with all relevant information about the method for realizing the scale and the instructions for the correct use of the CRM.

In certain cases the user can apply pure chemical compounds for realizing the fixed points if CRMs certified in the scale units are unavailable or expensive, or if their use is not necessary at the level of the uncertainty of the measurement. If this method is chosen, the user should be aware of the correlation between the purity of the material and the property on which the scale is based, and the uncertainty of the measurement can be only roughly estimated.

There is a great variety of conventional scales and the methods of application of the RMs for realizing them differ widely so that only some aspects can be discussed. Examples are given below to illustrate different features of some conventional scales.

3.2 The International Practical Temperature Scale

The International Practical Temperature Scale (IPIS-68) is based on assigned values of the temperature of 13 reproducible equilibrium states (defining fixed points) and on standard instruments calibrated at those temperatures. Interpolation between the fixed-point temperatures is provided by formulae used to establish the relation between indications of the standard instruments and values of International Practical Temperature. In addition there are 31 secondary reference points (i.e. equilibrium states of RMs).

The unit of thermodynamic temperature and of IPTS-68, the kelvin, is the fraction 1/273,16 of the thermodynamic temperature of the triple point of water. This temperature is unique in that it is exact by definition and its realization is reproducible within a range of 0,2 mK. At present it is not possible to assign levels of accuracy to the temperatures of the secondary reference points, except for those based on triple points.

A number of recommendations for realizing the reference points are published by the International Union of Pure and Applied Chemistry[6]. Some metrological institutes supply CRMs certifying the reference temperatures, but most of the RMs are supplied by reagent manufacturers and are characterized by purity. To measure the purity, cryoscopic methods, vapour pressure measurement, and gas chromatography are recommended.

3.3 The pH scale

Since absolute single ion activities cannot be measured experimentally, it is recognized that the pH value is an inexact physical quantity. In order that measured pH is endowed with as much significance as possible, a conventional pH scale has been adopted which is defined by reference solutions with assigned values of pH. These values have been determined by measuring the e.m.f. of a hydrogen – silver chloride cell without transference and by a given method of calculation, based on a convention.

Various national standard specifications describe the methods of preparing and assigning values of pH to the reference solutions. The requirement concerning the purity of materials is the analytical reagent grade. The uncertainty of the certified values should be limited to ± (0,003 to 0,010) pH unit.

3.4 The octane-number scale

The octane-number scale is defined by ASTM and IP joint standard specifications. The ISO standards[7,8], as well as a number of national standards, refer to ASTM-IP documents. ASTM D 2699-84 and ASTM D 2700-84 describe the test methods for knock characteristics of motor fuels by the research method and by the motor method respectively. In both standards the octane number of a fuel is determined by comparing its knocking tendency with those for blends of ASTM reference fuels of known octane number under standard operating conditions. The reference materials and blending accessories are given in annexes of both standards.

The ASTM standards refer to NBS SRM No. 1816 (iso-octane, purity 99,98 %) and SRM No. 1815 (*n*-heptane, purity 99,87 %). The principal use of these materials is in certifying the commercially produced ASTM Knock Test Reference Fuels. Specifications for these reference fuels are given in the standard, in which the suppliers are also listed. The responsibility for meeting the specifications for the reference materials rests with the suppliers. ASTM certification is based on the physical properties of the sample. Suppliers must test a sample of the reference material to be certified and at the same time test the corresponding SRM to provide traceability of production to a standard material. A certificate is issued by ASTM to the suppliers authorizing them to guarantee that the material shipped is so tested and to quote the results of the tests.

3.5 Soluble solids content by the refractive index method

ISO 2173[9], in accordance with other standards, specifies a refractometric method for the examination of fruit and vegetable products. For convenience the result obtained by this method may be considered conventionally as the soluble solids content. The soluble solids content is defined as the concentration of sucrose in an aqueous solution which has the same refractive index as the product analysed, under specified conditions of preparation and temperature. This concentration is expressed as a percentage by mass.

The recommended measuring instrument is a refractometer indicating the percentage by mass of sucrose. It must be adjusted to read a soluble solids (sucrose) content of zero for distilled water at 20 °C.

A refractometer adjusted to read a refractive index of 1,333 0 for distilled water at 20 °C can also be used. The tables correlating refractive index with soluble solids content (expressed as sucrose) are given in the standard specifications.

3.6 Unit of turbidity

The unit of turbidity is defined in ISO 7027[10]. The turbidity can be measured in formazine attenuation units (FAU) if the apparatus measures attenuated radiation, or in formazine nephelometric units (FNU) if diffused radiation is measured. The unit can be reproduced by preparing a given stock solution, the turbidity of which is 400 in FAU or in FNU. The stock solution may be diluted to obtain standard matching solutions of turbidities in the range of interest.

3.7 Hardness scales

The hardness of a material is an important practical, technological property. The Rockwell, Brinell, Vickers and Knoop scales, widely used in industry, are conventional scales. The means for transferring the unit from the standard apparatus to users' instruments are not regarded as reference materials, but as material measures. These measures (standardized blocks) should be verified according to OIML I.R. 9, 10, 11 and 12. Studying their role is beyond the scope of this Guide.

There are some other standardized methods (e.g. Martens, Shore, Jones, Sward, Hannemann) for hardness measurement; these are mostly related to the apparatus employed and are not based on reference materials.

3.8 Caking power of coals by the Roga method

The Roga test is widely used for the classification of hard coals. The test is standardized in many countries and is also described in ISO 335[11]. The measurement method and the necessary measuring accessories are strictly laid down. The Roga test is based on a standard reference material, which is an anthracite of defined physical properties and composition. Moreover, some standard specifications, e.g. the Hungarian, refer to a given reference anthracite.

The feature of the determination is that the reference material and the sample to be measured are first mixed and caked together. The Roga index can be calculated from the rest of the material after the measuring process.

Annex A

Bibliography

[1] ISO Guide 35 : 1985, *Certification of reference materials — General and statistical principles.*

[2] BIPM/IEC/ISO/OIML. *International vocabulary of basic and general terms in metrology* [*VIM*]. 1984.

[3] ISO Guide 30 : 1981, *Terms and definitions used in connection with reference materials.*

[4] ISO 3534-1 : —, *Statistics — Vocabulary and symbols — Part 1 : Probability and general statistical terms.* (Revision of ISO 3534 : 1977. To be published.)

[5] ISO 5725 : 1986, *Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests.*

[6] IUPAC. *Recommended reference materials for the realization of physicochemical properties.* K.N. Marsh (editor). Blackwell Scientific Publications, Oxford, 1987.

[7] ISO 5163 : 1977, *Motor and aviation-type fuels — Determination of knock characteristics — Motor method.*

[8] ISO 5164 : 1977, *Motor fuels — Determination of knock characteristics — Research method.*

[9] ISO 2173 : 1978, *Fruit and vegetable products — Determination of soluble solids content — Refractometric method.*

[10] ISO 7027 : 1984, *Water quality — Determination of turbidity.*

[11] ISO 335 : 1974, *Hard coal — Determination of caking power — Roga test.*

Further reading

ISO Guide 31 : 1981, *Contents of certificates of reference materials.*

ISO Handbook 3 : 1981, *Statistical methods.*