GUIDE 35

Certification of reference materials — General and statistical principles

Second edition 1989

Contents

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 35 was drawn up by the ISO Committee on reference materials (REMCO) and was submitted directly to ISO Council for acceptance. This second edition cancels and replaces the first edition (ISO Guide 35 : 1985), to which a new clause 9 has been added.

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

Certification of reference materials — General and statistical principles

Introduction

The Committee on reference materials (REMCO) is concerned with guidelines for the preparation, certification and use of reference materials. This Guide is intended to describe the general and statistical principles for the certification of reference materials.

Various sections of this Guide were prepared by different delegates to REMCO. The project was co-ordinated with representatives of ISO/TC 69, *Applications of statistical methods.*

Acknowledgment is given to J. D. Cox (BSI, UK) for preparation of the section on the role of reference materials in measurement systems (clause 3). Much of clauses 4, 5 and 6 is based on material contained in three previously published sources:

a) CALI, J.P. *et al.* The role of standard reference materials in measurement systems, *NBS Monograph 148,* Washington, DC, National Bureau of Standards, 1975 (especially Chapter III, by H. H. Ku);

b) URIANO, G. A. and GRAVATT, C. C. The role of reference materials and reference methods in chemical analysis. *Crit. Rev. in Anal. Chem.* **6** 1977 : 361;

c) MARSCHAL, A. *Matériaux de référence.* Bureau National de Métrologie, Laboratoire National d'Essais, Paris.

K. R. Eberhardt (ANSI, USA) prepared clause 7 on the use of a definitive method to certify reference materials. R. Sutarno and H. Steger (SCC, Canada) prepared clause 8 on the use of an interlaboratory testing programme to certify reference materials. H. Marchandise (Community Bureau of Reference, Commission of the European Communities) prepared clause 9 on a metrological approach to certification, included for the first time in the second edition of this Guide. G. Uriano (ANSI, USA) served as editor of the Guide.

Special acknowledgement is given to members of ISO/TC 69/SC 6 and its Secretary K. Petrick (DIN, Germany, F.R.), for their co-operation in preparing those sections of the document concerned with the statistical analysis of data. In particular the many contributions of Prof. P. T. Wilrich (DIN, Germany, F.R.) and Dr. T. Miyazu (JISC, Japan) of ISO/TC 69/SC 6 to the review and editing of the Guide are gratefully acknowledged.

Earlier Guides [1-3] prepared by REMCO have dealt with the following aspects of reference materials:

a) mention of reference materials in International Standards;

b) terms and definitions used in connection with reference materials;

c) the contents of certificates of reference materials.

The purpose of this Guide is to provide a basic introduction to concepts and practical aspects related to the certification of reference materials. ISO Guide 33 [29] more fully addresses concepts and practical aspects related to the use of reference materials.

1 Scope

According to the definition given in 2.1, reference materials (RMs) may be used in diverse measurement roles connected with instrument calibration, method assessment and assignment of property values. The purpose of clause 3 is to discuss these measurement roles and to show how traceability [1] of measurement may be secured by use of RMs, thus yielding worldwide compatibility of measurement.

Just as certified reference materials (CRMs) are to be preferred over other classes of RMs in citations in International Standards [1], so also are CRMs to be preferred over other classes of RMs in measurement science generally, given that CRMs needed for a particular type of measurement exist. Assistance in locating the source(s) of supply of CRMs for various technical fields is afforded by ISO's *Directory of certified reference materials* [4].

It will be evident that the quality of a measurement based on use of a CRM will depend in part on the effort and care expended by the certifying body on determining the property

1) An internationally agreed definition of "traceability" in measurement science is given in reference [5]:

traceability: The property of a result of a measurement whereby it can be related to appropriate standards, generally international or national standards, through an unbroken chain of comparisons.

value(s) of the candidate CRM. Hence the process of certification [2] should be carried out using well-characterized measurement methods that have high accuracy as well as precision and provide property values traceable to fundamental units of measurement. Furthermore, the methods should yield values with uncertainties that are appropriate to the expected end-use of the CRM. Clauses 4 and 5 deal with two of the most important technical considerations in the certification of RMs — measurement uncertainties and material homogeneity. Clause 6 provides general principles for RM certification.

Two commonly used general approaches to assuring technically valid RM certification are discussed in clauses 7 and 8. Clause 7 describes the use of a single method of the highest accuracy (i.e. sometimes referred to as a "definitive" or "absolute" method) and usually employed by a single laboratory for RM certification. Clause 8 describes the use of an interlaboratory testing approach to RM certification, which might involve more than one method.

The metrological approach discussed in clause 9 has as its objective the production of certified values the accuracy and uncertainty of which are demonstrated by experimental evidence.

In summary, the purpose of this Guide is to assist in understanding valid methods for the certification of RMs and also to help potential users to better define their technical requirements. The Guide should be useful in establishing the full potential of CRMs as aids to assuring the accuracy and interlaboratory compatibility of measurements on a national or international scale.

2 Definitions

Definitions of the basic terms "reference material" and "certified reference material" were first put forward in 1977 [1] and were later amended slightly [2] to read as follows.

2.1 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.

NOTE — An RM may be in the form of a pure or mixed gas, liquid or solid, or even a simple manufactured object. Some RMs are certified in a batch, any reasonably small part of which should exhibit the property value(s) established for the whole batch within stated uncertainty limits. Other RMs exist as individually manufactured objects which are also certified individually. Numerous RMs have properties which, because they cannot be correlated with an established chemical structure or for other reasons, cannot be measured in mass or amount of substance units or determined by exactly defined physical or chemical measurement methods. Such RMs include certain biological RMs (for example a vaccine to which an international unit has been assigned by the World Health Organization) and certain technological RMs (for example rubber blocks for the determination of abrasiveness or steel plates for the determination of hardness). It is recognized that the definition of "reference material" given above could involve an overlap with the term "material measure" as defined in the *International Vocabulary of Basic and General terms in Metrology* [5]; consequently, some materials may be characterized as either reference materials or material measures.

2.2 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.

NOTE — A CRM may consist of units which are each certified individually or which are certified by examination of representative samples from a batch.

3 The role of reference materials in measurement science

Metrology is the field of knowledge concerned with measurement. Metrology or measurement science [1] includes all aspects both theoretical and practical with reference to measurements, whatever their level of accuracy, and in whatever fields of science or technology they occur [6]. This clause describes the role of reference materials in quantitative measurements.

3.1 The role of reference materials in the storage and transfer of information or property values

By definition (2.1), a reference material has one or more properties, the values of which are well established by measurement. Once the property value(s) of a particular RM have been established, they are "stored" by the RM (up to its expiration date) and are transferred when the RM itself is conveyed from one place to another. To the extent that the property value of an RM can be determined with a well-defined uncertainty, that property value can be used as a reference value for intercomparison or transfer purposes. Hence RMs aid in measurement transfer, in time and space, similar to measuring instruments [2] and material measures [6].

A general scheme for constructing a hierarchical measurement system is illustrated in section 6.5 of the *Vocabulary of Legal Metrology* [6]. The interlinking of various levels and stations within a measurement system via "reference standards" may, in principle, be effected by either measuring instruments or material measures or RMs.

An RM must be suitable for the exacting role it performs in storing and transferring information on measured property values. The following technical criteria (legal or commercial criteria

1) "Measurement science" is therefore synonymous with "metrology" according to the international definition of the latter term [6]; it should be noted, however, that current usage generally restricts the term "metrology" to physical measurements at high accuracy. The term "metrology" is, however, being increasingly used in the context of chemical, engineering, biological and medical measurements.

2) Some measuring instruments are not readily movable (by reason of size, mass, fragility, instability or cost), in which case the measurand must be brought to the instrument to effect the measurement transfer. But all RMs and material measures are readily movable and thus can be taken to the measurand.

may be relevant also) apply to the fitness for purpose of RMs in general :

a) the RM itself and the property value(s) embodied in it should be stable for an acceptable time-span, under realistic conditions of storage, transport and use;

b) the RM should be sufficiently homogeneous that the property value(s) measured on one portion of the batch should apply to any other portion of the batch within acceptable limits of uncertainty; in cases of inhomogeneity of the large batch, it may be necessary to certify each unit from the batch separately;

c) the property value(s) of the RM should have been established with a precision and an accuracy sufficient to the end use(s) of the RM;

d) clear documentation concerning the RM and its established property value(s) should be available. Preferably the property value(s) should have been certified, so the documentation should then include a certificate, prepared in accordance with ISO Guide 31 [3].

The word "accuracy" was advisedly used in c) to indicate that whenever possible, the measurement of a given property value should have been made by a method having negligible systematic error or bias relative to end-use requirements (or where the result has been corrected for a known bias) and by means of measuring instruments or material measures which are traceable to national measurement standards. Subsequent use of an RM with traceable property values ensures that traceability is propagated to the user. Since most national measurement standards are themselves harmonized internationally, it follows that measurement standards in one country should be compatible with similar measurements in another country. In many cases, CRMs are appropriate for the intercomparisons of national measurement standards.

3.2 The role of reference materials in the International System of units (SI)

3.2.1 Dependence of the SI base units on substances and materials

The majority of measurements made in the world today are within the framework of the International System of units [7]. In its present form, SI recognizes seven base units, namely the units of length (metre, symbol m), mass (kilogram, kg), time (second, s), electric current (ampere, A), thermodynamic temperature (kelvin, K), amount of substance (mole, mol) and luminous intensity (candela, cd). The definitions [7] of these base units mention the following substances : krypton-86 [1] (for defining the metre), platinum-iridium (for fabricating the prototype kilogram), caesium-133 (for defining the second), water (for defining the kelvin) and carbon-12 (for defining the mole). Opinions differ as to whether the substances named fall under the definition of reference material (2.1). The use of these substances in basic metrology is consistent with the use of reference materials in other types of measurement applications. Certainly such materials have a special status as defined substances on which the SI is based. The dependency strictly applies to definition of the unit, since realization of the units may involve other substances/materials. This is especially true in regard to the realization of the mole [8] and the kilogram.

3.2.2 The realization of derived SI units with the aid of reference materials

From the seven base units an unlimited number of derived units of the SI are obtainable by combining base units as products and/or quotients. For example, a derived unit of mass concentration is defined as $kg \cdot m^{-3}$ and the derived unit of pressure (given the special name pascal, symbol Pa) is defined as $m^{-1} \cdot kg \cdot s^{-2}$. Formally speaking, the derived units ultimately depend on the substances on which the base units themselves depend (see 3.2.1). In practice, the derived units are often realized not from base units but from RMs with accepted property values. Thus a variety of substances/materials may be involved in the realization of derived units (examples 1 and 2 below) or even of base units (examples 3 and 4 below).

Example 1: The SI unit of dynamic viscosity, the pascal second ($Pa \cdot s = m^{-1} \cdot kg \cdot s^{-1}$) may be realized [9] by taking the value for a well purified sample of water as 0,001 002 Pa·s at 20 °C.

Example 2: The SI unit of molar heat capacity, the joule per mole·kelvin ($J \cdot mol^{-1} \cdot K^{-1} = kg \cdot m^2 \cdot s^{-2} \cdot mol^{-1} \cdot K^{-1}$) may be realized [10] by taking the value for purified α-alumina as 79,01 $J \cdot mol^{-1} \cdot K^{-1}$ at 25 °C.

Example 3: The SI unit of amount of substance, the mole, may be realized by taking 0,069 72 kg of highly purified gallium metal [11].

Example 4: The SI unit of temperature, the kelvin, may be realized at any temperature T_1 (273,15 K < T_1 < 903,89 K) from measurements of the resistance of a highly pure platinum wire at T_1, at the triple point of purified water, at the freezing point of purified tin and at the freezing point of purified zinc, coupled with use of a specified mathematical relation [12]. The word "thermodynamic" has been deliberately omitted here to avoid controversy over whether thermodynamic temperatures are, or are not, the same as International Practical Temperatures of 1968 : the intention of the International Committee for Weights and Measures was to match the two sorts of temperature exactly, within the framework of knowledge available during 1968-1975.

3.2.3 Connection of analytical chemistry to the International System of units

It will be noted that purified (often called "pure") chemical substances were cited in each of the examples 1 to 4 (3.2.2). The measurement of degree of purity, or more generally of the chemical composition of materials, is within the realm of analytical chemistry. In addition to the dependence of SI on chemical substances, the dependence of analytical chemistry on SI is worthy of examination. Presently, most analytical

1) Recently, the General Conference on Weights and Measures redefined the metre as the distance travelled by light in a vacuum during 1/299 792 458 of a second.

chemists employ units within the SI (all base units except the candela and also many derived units) in their measurements. However, compositional analysis depends on an additional concept, namely that pure chemical species exist to which the chemical compositions of other substances and materials are referred, by invoking the laws of chemical change and stoichiometry.

From one or more pure chemical species, considered to be primary measurement standards, it is feasible to construct measurement hierarchies for analytical chemistry similar to those used in physical measurement [6]. Examples of such measurement standards are:

a) the electron, to which other species can be connected by electrochemical analysis [13];

b) carbon-12, to which other species can in principle be connected by mass spectrometry, Raoult's law measurements, or volumetric measurements with low-density gases, etc.;

c) a highly purified element or compound, to which other species can be connected by electrochemical, gravimetric, titrimetric, spectrometric methods, etc.

The "other species" cited in these examples will in many cases be used as RMs. Many substances can fill this role of intermediaries between primary and working analytical standards using the diversity of techniques and chemical reactions that an analyst may employ. The concept of traceability applies to analytical chemistry as much as it does to other branches of measurement science. The quality of the result of a chemical analysis will be enhanced if the result's traceability can be clearly stated in terms of the traceability of the instruments, material measures and RMs employed. In most cases, the traceability will also depend on the values of the relative atomic masses (formerly called "atomic weights") used in the calculations; the source of these should be recorded by the analyst (for example [11]).

3.2.4 The role of reference materials in realizing units outside of the SI

Where the components of a measurement system (for example the Imperial system) can be related exactly to the corresponding components of the SI, it is unnecessary to have independent means for realizing the non-SI measurement system. Where the quantities cannot be related to those of the SI, then independent realization of the non-SI units is in principle necessary. In practice, however, few such systems remain in use and thus are mostly historical curiosities.

3.3 Use of reference materials

REMCO intends to publish a separate guide covering general and statistical principles for the use of reference materials. There are very few published documents that address general problems associated with the use of reference materials. The reader is referred to the documents and recommendations published by IUPAC Commission I.4 on Physico-chemical Reference Materials and Standards, which deal primarily with the use of reference materials for realization of physical properties. The following IUPAC Commission I.4 publications in *Pure and Applied Chemistry* are concerned with the certification and use of reference materials for physical properties:

Physical property	Volume, date of publication and page number
Enthalpy	**40** 1974 : 399
Optical rotation	**40** 1974 : 451
Optical refraction	**40** 1974 : 463
Density	**45** 1976 : 1
Relative molecular mass	**48** 1976 : 241
Absorbance and wavelength	**49** 1977 : 661
Reflectance	**50** 1978 : 1 477
Potentiometric ion activities	**50** 1978 : 1 485
Viscosity	**52** 1980 : 2 393
Permittivity	**53** 1981 : 1 847
Thermal conductivity	**53** 1981 : 1 863

4 Measurement uncertainty

In discussing measurement uncertainties, the terms "precision", "systematic error or bias", and "accuracy" are usually used. The meanings of these terms are not rigidly fixed, but depend to a large extent on the interpretation and use of the data [14, 15].

4.1 An illustrative example

If two equally trained operators, A and B, each make four replications of a measurement on a uniform material each day for 4 days on one instrument, and 4 days again on a similar instrument, the results, 16 sets of four measurements, may look like those in figure 1. What can be seen from this plot ?

a) the spreads among each set of four values are comparable, perhaps slightly smaller for instrument 2 than instrument 1;

b) there appears to be more variability between daily results than within sets of daily results, particularly for instrument 1;

c) operator B gives lower results than operator A;

d) instrument 1 gives lower results than instrument 2.

Figure 1 is constructed for the purpose of demonstration, and actual measurements could be better or worse than shown. However, this plot does show some four types of factors that contributed to the total variability of these measurements:

1) factors acting within days;

2) factors acting between days;

3) factors due to instrument systems;

4) factors due to operators.

Appropriate techniques are available for the separate estimation of the effects of these four factors and standard deviations could be computed corresponding to each of them. However, the limited number of operators and instruments prevents the computation of standard deviations as reliably for

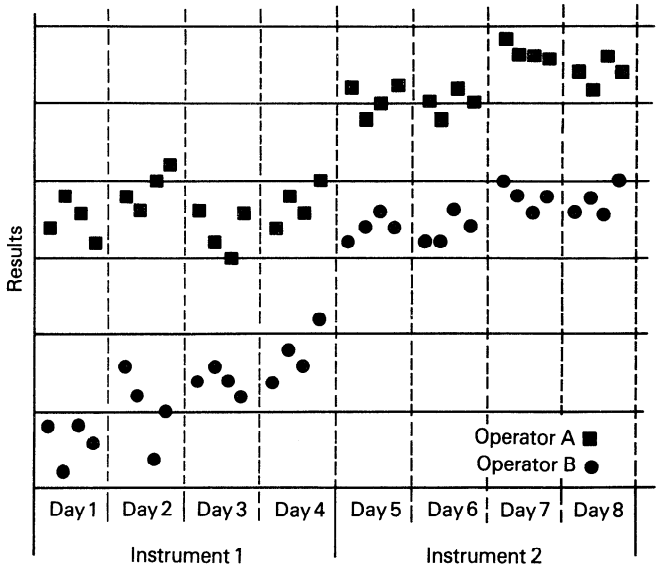


Figure 1 — An example of results of measurements by two operators using two instruments on eight different days

factors 3) and 4) as for factors 1) and 2). The time and work involved certainly impose limits on any efforts to do so.

The failure to allow for factors relating to instruments and operators is one of the main causes for the unreasonable differences usually encountered in interlaboratory, or round-robin, types of tests [16]. Because instruments vary from time to time and operators change, the result from a laboratory at a given time represents only one of the many results that could be obtained, and the variability caused by these two sources must be considered as part of the precision of the laboratory. The standard deviation computed without regard to these effects would underestimate the true variability.

If, by the proper use of standards and reference methods [17], these two sources of errors were eliminated, the standard deviation computed from the 16 means of sets of four measurements would be the proper measure of precision. Presumably the grand mean of the 16 mean values would be reported.

The mean of many values is more stable than individual measurements. When extraneous sources of variation, such as instrument and operator effects, are eliminated, the relationship between the standard deviation of individual measurements and the standard deviation of the mean of n such measurements can be expressed as

$$\sigma(\bar{X}_n) = \frac{\sigma(X)}{\sqrt{n}} \qquad \ldots (1)$$

In other words, the standard deviation of the mean is smaller than the standard deviation of individual measurements by a factor of $1/\sqrt{n}$. One important provision must hold for this relationship to be true, i.e. that the n measurements are independent of each other. "Independence" can be defined in a probability sense, but for present purposes, measurements may be considered independent if they show no trend or pattern. This is certainly not true in figure 1, and to say that the standard deviation of the mean of all 64 values is 1/8 ($= 1/\sqrt{64}$) of the standard deviation of individual measurements would seriously underestimate its true variability. Moreover, the relationship in equation (1) is expressed in terms of the true value of the standard deviation, σ, which is usually not known. As the computed standard deviation, s, is itself an estimate of σ from the set of measured values, the standard deviation of the mean in equation (1) is only approximated when s is used in place of σ.

The use of the standard deviation computed from daily averages rather than individual values is preferred because the former properly reflects a component of variability between days, or over time, which is usually present in precision measurement.

4.2 Some basic statistical concepts

The basic information available on the measurement errors is summarized by:

a) the number of independent determinations or the number from which a mean was computed and reported;

b) an estimate of the standard deviation, s, defined by

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$$

where n measurement results are denoted by $x_1, x_2, \ldots, x_n$, and their mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

From a) and b) several useful derived statistics can be computed:

c) standard deviation of the mean of n measurements

$$s(\bar{x}_n) = \frac{s}{\sqrt{n}}$$

This is sometimes called the standard error of the mean to differentiate it from the standard deviation of individual determinations.

NOTE — As n becomes large, the value of $s(\bar{x}_n)$ becomes very small, showing that the average of a large number of measurements approaches a constant value μ which is usually the objective of the measurement procedure.

d) confidence interval for the mean (normal distribution). Each time n measurements are made, a value of the average of the measurements is reported. These averages will differ from time to time within certain limits. Assuming a normal distribution, one interval of the type $\bar{x} \pm \delta$ can be constructed [18] such that the interval from $\bar{x} - \delta$ to $\bar{x} + \delta$ will

be fairly certain to include the value of μ desired. The interval is computed by:

$$\delta = t \frac{s}{\sqrt{n}} \quad \ldots (2)$$

where t is a tabular value of the Student distribution, and depends on the confidence level and the degrees of freedom for s;

e) 2-sigma (or $2s$), 3-sigma (or $3s$) limits. These limits describe the distribution of measurement error. If a measurement is made by the user of a CRM having the same precision (i.e. same σ) as that obtained by the certifying laboratory, his measurements should fall (with probability approximately 0,95 to 0,997) within these limits when σ is well-established. Otherwise there is evidence of systematic difference.

4.3 Instrument and operator errors

Instrument and operator types of errors have not yet been treated. An ideal situation would be to eliminate them from the measurement process, or to use more instruments and more operators and then estimate standard deviations associated with these sources. When neither of the above is feasible or practical, the least that can be done is to use two instruments and/or operators. If the confidence intervals for the mean results of the two instruments do not overlap, then there is good evidence of instrument difference.

Using his experience and judgement, a measurement scientist may arrive at reasonable bounds for these types of errors. If the bound is not computed from measurement data, then its validity cannot be supported by statistical analysis. In such cases, these bounds are "guesstimates" and the only recourse is to treat them as limits to systematic errors.

The detection of differences and the separation of the total variability into its identifiable components can be facilitated through careful planning and statistical design of the experiment.

4.4 Differences among measurement methods

Each measurement method purports to measure the desired property of a material, but seldom does a method measure the property directly. In most cases the method actually measures some other property that is related to the property by theory, practice, or tradition, and then converted to the value of the desired property through these relationships. Discrepancies among results of different measurement methods are common, even for measurements leading to the determination of fundamental physical constants [19].

In the preparation of a CRM, usually two or more measurement methods are employed for each property measured. If these methods are well established by virtue of past experience, the results yielded by these methods usually agree to within the uncertainty assigned to each method.

In a few cases these differences are so large that the results cannot be reconciled, and these results are then reported separately for each individual method. The RM is either not certified or certified on a method-dependent basis. A historical example of this type of reporting is NBS CRM 1091, Stainless Steel. The nitrogen content was measured by vacuum fusion and pressure bomb-distillation, and gave results of 861 and 945 mg/kg, with standard deviations of 3 and 20 mg/kg, respectively. Clearly one or both methods have a systematic error that is large compared to the variability of material or the measurement uncertainty. A report of the average of the two methods would be highly misleading.

Measurement accuracy in its absolute sense is never realized. In practice, certified values of some reference materials are defined by using a referee method or assigning a value by a well-defined procedure so that at least the same benchmark will be used by everyone in the field. The importance of reference methods to supplement the use of these measurement standards is also being emphasized [17]. A good example is the reference method for blood haemoglobin and the value assigned as a benchmark to the reference material issued by the International Committee for Standardization in Hematology (ICSH) [20, 21].

4.5 Uncertainties of certified values

The uncertainty of a CRM value is usually made up of several components, some supported by data and some not:

a) a statistical tolerance interval giving bounds to material inhomogeneity based on data and statistical computations;

b) a confidence interval for the mean giving bounds to measurement error based on data and statistical computations;

c) components of measurement uncertainty due to variation among laboratories and/or operators and measurement methods;

d) a combination (addition of absolute values or the square root of the sum of the squares) of estimated bounds to "known" sources of possible systematic error based on experience and judgement (in other words, there are no data, or an insufficient number of data, to make a statistical calculation).

The word "known" is quoted above to contrast with systematic errors that are "unknown" or unsuspected. These unsuspected errors could occur in a number of ways — a component in the physical system, a minor flaw in the theoretical consideration, or the rounding error in a computation. As more homogeneous materials become available, and more precise measurement methods are developed, these types of errors will be detected by design or by chance and hopefully will be eliminated. Improved accuracy in the measurement of a property is basically an expensive iterative process and unwarranted demand for accuracy could mean the waste of resources.

4.6 Statements of uncertainty on CRM certificates

A variety of statements of uncertainty can be found in past and current certificates issued for CRMs around the world. Some of these statements are well formulated and supported by data,

others are not; some of these statements contain a wealth of information that is useful to exacting users, but overwhelming to others; some statements are oversimplified with a resulting loss of information. Because the originator of a CRM has to keep all classes of users in mind, the use of a single form of statement is not usually possible. The intention is that all these statements are unambiguous, meaningful, and contain all the information that is relevant for potential users.

Some commonly used statements, taken from existing certificates, are listed in 4.6.1 to 4.6.4.

4.6.1 Example 1: 95 % confidence limits for the mean

Rubidium chloride

Absolute abundance ratio ..... 2,593 ± 0,002

"The indicated uncertainties are overall limits of error based on 95 % confidence limits for the mean and allowances for the effects of known sources of possible systematic error."

Because the isotopic ratio is a constant for a given batch of material and is not subject to errors of material inhomogeneity, the 95 % confidence limits for the mean refer to measurement error only. This is computed from

$$t \frac{s}{\sqrt{n}}$$

as described in equation (2).

The effects of known sources of possible systematic error are discussed in detail in "Absolute isotopic abundance ratio and atomic weight of terrestial rubidium" [22].

4.6.2 Example 2: 2-sigma or 3-sigma limits

Glass Filters for Molecular Absorption Spectrometry

Absorbance ..... 0,500 0 ± 0,002 5

"This uncertainty is the sum of the random error of ± 0,1 % relative (2σ limit) and of estimated biases which are ± 0,4 % relative."

Each glass filter was individually calibrated, and the standard deviation refers to measurement error, including the cleanliness of the surface. As these glass filters will be used time after time, a multiple of the standard deviation is a proper measure of variability.

4.6.3 Example 3: Uncertainty expressed in significant digits

AISI 4340 Steel

Element Mass Fraction

Carbon ..... $3{,}8_2 \times 10^{-3}$

Manganese ..... $6{,}6 \times 10^{-3}$

According to the explanation given in the text: "The value listed is not expected to deviate from the true value by more than ± 1 in the last significant figure reported; for a subscript figure, the deviation is not expected to be more than ± 5." Thus, the mass fraction of carbon, expressed as a percentage, is between 0,377 and 0,387; and that for manganese is between 0,65 and 0,67. These uncertainties include material inhomogeneity, measurement imprecision, and possible bias between laboratories and implicit rounding, because these values are "... the present best estimate of the true value based on the results of a co-operative interlaboratory analytical programme."

When 20 to 30 elements are to be certified for one material, this method gives a concise and convenient summary of the results. As these limits are expressed in units of 5 and 10, some information is unavoidably lost for some of the elements. However, when the certified value is used, it is important to use all of the digits given including the subscripts. The uncertainty stated on this certificate depends heavily on the use of chemical judgement.

4.6.4 Example 4 : Standard deviation, and number of determinations

Method		Oxygen in ferrous metals (µg/g)		
		CRM A (Ingot iron)	CRM B (Stainless steel : AISI 431)	CRM C (Vacuum melted steel)
Vacuum fusion				
	$\bar{x}$	484	131	28
	s	14	8	2
	n	216	286	105
Neutron activation				
	$\bar{x}$	492	132	28
	s	28	7	4
	n	6	6	5
Inert gas fusion				
	$\bar{x}$	497	129	29
	s	13	8	5
	n	12	11	20

where

$\bar{x}$ is the mean oxygen value;

s is the standard deviation of an individual determination;

n is the number of determinations.

NOTE — The standard deviation includes error due both to the imprecision of the analytical method and to possible heterogeneity of the material analysed.

One criticism against this mode of presentation is that the user will have to compute the uncertainty based on his own understanding of the relationships.

5 Homogeneity of materials

Most RMs are subjected to a preparation procedure which ultimately includes subdivision into usable units. A subset of individual units from the batch is chosen for measurement according to a statistically valid sampling plan. A measurement uncertainty is derived taking into account material inhomogeneity as well as other factors (see clause 4). Other types of RM are prepared as individual artifacts and the certification is based on separate measurement of each unit rather than on statistical sampling of the complete batch. The second approach is useful when the RM can be measured non-destructively.

5.1 Materials

RMs prepared as solutions or pure compounds are expected to be homogeneous on physical (thermodynamic) grounds. The object of the test for homogeneity is mainly to detect any impurities, interferences or irregularities.

Materials such as mixed powders, ores, alloys, etc. are heterogeneous in composition by nature. RMs prepared from such materials must therefore be tested to assess the degree of homogeneity.

5.2 Concept of homogeneity

In theory, a material is perfectly homogeneous with respect to a given characteristic if there is no difference between the value of this characteristic from one part (unit) to another. However, in practice a material is accepted to be homogeneous with respect to a given characteristic if a difference between the value of this characteristic from one part (or unit) to another cannot be detected experimentally. The practical concept of homogeneity therefore embodies both a specificity to the characteristic and a parameter of measurement (usually the standard deviation) of the measurement method used, including the defined sample size of the test portion.

5.2.1 Characteristic of interest

A material may be sufficiently homogeneous with respect to the characteristic of interest to be useful as an RM even though it is inhomogeneous with respect to other characteristics, provided that this inhomogeneity exerts no detectable influence on the accuracy and precision of the commonly used methods of determination for the characteristic of interest.

5.2.2 Homogeneity measurement method

The degree of homogeneity that a material must have for use as an RM is commensurate with the precision attainable by the best available methods for the determination of the characteristic for which the RM is intended. Therefore, the greater the precision of the measurement method, the higher is the required degree of homogeneity of the material.

The precision attainable by the homogeneity measurement method varies with both the characteristic measured and its value for the RM. An RM intended for more than one characteristic is described by a corresponding number of statements of homogeneity, each of which should be traceable to an experimentally determined precision. The magnitude of the precision can vary widely.

In many cases, the precision attainable by a measurement method is affected by the size of the test portion taken from the RM. The degree of homogeneity of an RM is therefore defined for a given test portion size.

5.2.3 Practice

Ideally, an RM should be characterized with respect to the degree of homogeneity for each characteristic of interest. For RMs intended for a relatively large number of characteristics, the assessment of the degree of homogeneity for all characteristics is both economically and physically burdensome, and in some cases unfeasible. In practice therefore, the degree of homogeneity of such RMs is assessed only for selected characteristics. It is recommended that these characteristics be appropriately selected on the basis of established chemical or physical relationships; for example, an interelement concomitance in the mineral phases of an RM makes reasonable the assumption that the RM also has an acceptable degree of homogeneity for the non-selected elements.

5.3 Experimental design

5.3.1 Objectives

For reference materials that are expected to be homogeneous on physical grounds, the main purpose of homogeneity testing is to detect unexpected problems. Some examples are differential contamination during the final packaging into individual units, or incomplete dissolution or equilibration of an analyte in a solvent (which could lead to steadily changing concentrations from the first vial filled to the last). A statistical trend analysis would be helpful in the latter case. If the material is produced in more than one batch, it is necessary to test the equality of the batches (or to certify the batches separately).

When the nature of a reference material leads one to expect some inhomogeneity, the goal of the testing programme is not simply detection of inhomogeneity, but rather the estimation of its magnitude. This may require a more extensive testing programme than is required for detection.

Inhomogeneity can manifest itself in at least two ways :

a) different subsamples of an RM unit may differ on the property of interest;

b) there may be differences between units of the RM.

Differences among subsamples can usually be reduced or controlled to an acceptably low level by making the size of the subsample sufficiently large. Often a study to determine the appropriate subsample size is conducted before the certification experiments are begun. Differences which exist between individual units of the candidate RM must be reflected in the uncertainty statement on the certificate.

In statistical terms, the experimental design must satisfy the following objectives:

1) to detect whether the within-unit (short-range) variation is statistically significant in comparison with the known variation of the measurement method;

2) to detect whether the between-units (long-range) variation is statistically significant in comparison with the within-unit variation;

3) to conclude whether a detected statistical significance for one or both of the within-unit and between-units variations indicates a corresponding physical significance of sufficient magnitude to disqualify the candidate RM for the intended use.

The degree of homogeneity of a candidate RM in final form should be known. The task for the assessment of the homogeneity can, however, be performed in several steps.

5.3.2 Preliminary test for homogeneity

A preliminary assessment of the homogeneity of a candidate RM can be performed after homogenization as an integral part of the preparation process. The physical properties of an RM that can cause segregation to occur, for example the type of blender, strongly influence the manner of sample selection. The samples should be taken at regions where physical differences are expected to occur. Random sampling should be adopted only when causes of physical differences are unknown or believed to be absent.

The number of samples taken and replicate determinations thereon should be such that the appropriate statistical test should be capable of detecting the possible existence of inhomogeneity at a predetermined level.

NOTE — ASTM E 826-81, *Standard practice for testing homogeneity of materials for the development of reference materials*, gives one detailed procedure for testing homogeneity of bulk material. This standard practice is specialized to the case of testing the homogeneity of metals, in either solid or powdered form, and finely ground oxide materials that are intended for use as reference materials in X-ray emission, or optical emission spectroscopy, or both. For most RM certification programmes, an appropriate preliminary test for homogeneity can be obtained by straightforward adaptation of the practice given in ASTM E 826-81.

5.3.3 Principal test for homogeneity

This test must be performed for the candidate RM after it has been packaged into final form regardless of whether a preliminary test for homogeneity has been done. The purpose of the test is to confirm that the between-units variation is not statistically and practically significant.

The units should be selected from the stock at random to give each unit an equal chance for selection. An experimental design should be used in which k units of material are selected and n replicate determinations are performed for each unit. It is recommended that the determinations be performed in random order to avoid possible systematic time variations. k and n should be sufficiently large to detect the possible existence of inhomogeneity at a predetermined level.

For certain RMs, replicate within-unit determinations are not possible because the use of the entire unit is prescribed by the producer. In this instance, the between-units variance must be compared with the estimated precision of the measurement method to assess the degree of homogeneity of the RM.

5.4 Possible outcomes of homogeneity testing

The selection of samples and the analysis of data are usually performed in consultation with a statistician. Depending on the form of material, the emphasis may be to detect trends or patterns, for example from one end to the other of a steel rod, from the centre to the edge of a plate, from the top to the bottom portion of bulk material in a drum; or to check on the variability of material among ampoules or bottles. A proper, statistically designed experiment helps to assure that conclusions are valid, and minimizes the number of measurements needed to reach such conclusions.

The possible outcomes of homogeneity testing are described in 5.4.1 to 5.4.3.

5.4.1 Very homogeneous material

Homogeneity is not a problem, or material variability is negligible in relation to either measurement errors or to the use of the CRM. In this case, the certified value is the best estimate of the mean property value for the lot and the allowance for uncertainty describes possible measurement error associated with that estimate.

5.4.2 Very inhomogeneous material

Material variability is a major factor in the total uncertainty. In this case the entire lot of material is rejected or reworked, or each specimen is individually measured and certified.

Reworking is a reasonable course of action when there is reason to believe that the source of inhomogeneity can be eliminated by preparing a new batch of material using improved procedures. However, this is not always possible, and it is sometimes necessary to tolerate a small amount of between-units inhomogeneity when the material cannot practically be improved.

5.4.3 Material of moderate homogeneity

Material variability is of the same magnitude as the measurement error, and must be included as a component of the uncertainty. This case is discussed in 5.5.

5.5 Some examples of homogeneity testing

Of the three cases (5.4.1 to 5.4.3) the last is the one most frequently encountered. Two subclasses are apparent : one where a trend is detected and one where no trend is detected.

Where a trend has been detected, for example along a steel rod to be cut into pieces, the unusable portion is discarded and, hopefully, the trend in the remaining portion is linear or can otherwise be described mathematically. In such cases, a line (or other appropriate mathematical expression) can be fitted to the values measured along the rod. The maximum departure from the average points on the fitted line is taken as a measure of inhomogeneity, assuming measurement error is small in comparison to the trend.

Where no trend is detected, but the results of measurements show variability that is not negligible, a statistical concept called

"statistical tolerance interval" can be used. To illustrate this concept, suppose a solution is prepared and packaged into 1 000 ampoules, of which 30 are measured for some property. For this example, the tolerance limit concept [18] states essentially that based on the measured values of the 30 ampoules almost all of the 1 000 ampoules will not differ from the average of the 30 ampoules by more than the constructed limit. In statistical terms, it would read: "The tolerance interval (mean $\pm$ Δ) is constructed such that it will cover at least 95 % of the population with probability 0,99". [1)]

This statement does not guarantee that the tolerance interval will include all of the ampoules. It says that 99 % of the time the tolerance interval will include at least 95 % of the ampoules. The "99 % of the time" refers to the way this tolerance interval is constructed, i.e., if 30 ampoules were selected from the population repeatedly, and the same experiments were performed over and over again, 99 % of the tolerance intervals so constructed would cover at least the proportion (95 %) of the total population as specified, and 1 % of the tolerance intervals would cover less than 95 % of the total population.

How is this interval constructed ? First, the mean [equation (3)] and standard deviation [equation (4)] from the 30 ampoules are computed :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \ldots (3)$$

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} \qquad \ldots (4)$$

where

$x_1, x_2, \ldots, x_i, \ldots, x_n$ are the measured values, with $n = 30$;

$\bar{x}$ is an estimate of the mean, μ, of the 1 000 ampoules;

s is an estimate of the measure of the dispersion, σ, among these ampoules.

The values $\bar{x}$ and s contain practically all the information available on the 1 000 ampoules and can be used to calculate the tolerance interval $\bar{x} \pm \Delta$.

The value of Δ is computed as a multiple of s, i.e. $\Delta = k'_2 s$. The value of k'_2 depends on three parameters :

a) the number, n, of samples measured (30);

b) the proportion, p, of the total population to be covered (0,95);

c) the probability level, $1 - \alpha$, specified (0,99).

A table of factors for two-sided tolerance limits for normal distributions gives the value for k'_2 as 2,841 for $n = 30$; $1 - \alpha = 0{,}99$; and $p = 0{,}95$. Tables of these factors are given in ISO 3207 [2)] and in many standard statistical texts [18].

The term "two-sided" means that we are interested in both over and under limits from the average. The term "normal distribution" refers to the distribution of all the values of interest and is a symmetrical, bell-shaped distribution usually encountered in precision measurement work.

Figure 2 is a histogram of the ratios of the emission rate of ^{137}Cs, in a ^{137}Cs nuclear fuel burn-up reference material, to a radium reference standard. A frequency curve of a normal distribution can be fitted to these data. There were 98 ampoules of ^{137}Cs involved; each ampoule was measured in April, September, and November, 1972. By averaging the three measurements, the measurement error was considerably smaller than the difference of masses of active solutions among these ampoules, and the plot in figure 2 shows essentially the inhomogeneity of the mass of solution in the ampoules.

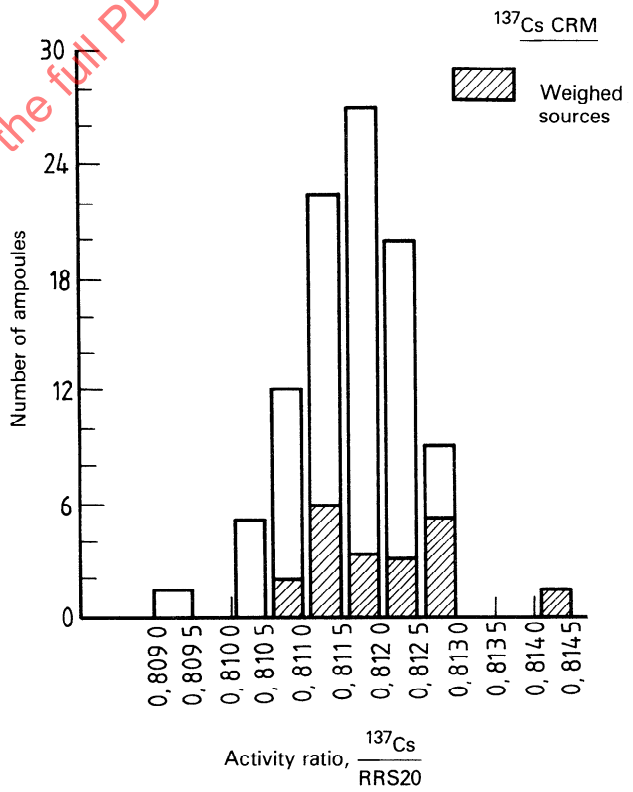


Figure 2 — Histogram of the frequency (number of ampoules) versus the ratio of the activity of ^{137}Cs standards to a radium reference standard (RRS20)

1) The statement is true only for a population of infinite size; however, the correction for a population of finite size is negligible where finite size is large.

2) ISO 3207, *Statistical interpretation of data — Determination of a statistical tolerance interval.*

To illustrate the concept of tolerance interval properties, values of 30 ampoules were selected from the 98 values by using a random number table, and $\bar{x}$ and s were computed by equations (3) and (4), giving

$$\bar{x} = 0{,}811\ 68 \qquad s = 0{,}000\ 80$$

The tolerance interval for at least 95 % coverage with probability level 0,99 is $\bar{x} \pm k'_2 s$, or

$$0{,}811\ 68 - (2{,}841 \times 0{,}000\ 80) = 0{,}809\ 41$$

to

$$0{,}811\ 68 + (2{,}841 \times 0{,}000\ 80) = 0{,}813\ 95$$

This interval covers the values of 96 ampoules, i.e. all except the single ampoule on the extreme left and the one on the extreme right.

If 30 values were selected repeatedly, the values of $\bar{x}$ and s would not be the same, but the intervals $\bar{x} - k'_2 s$ to $\bar{x} + k'_2 s$ would have the property of covering at least 95 % of all values 99 % of the time.

Tolerance limits and intervals are useful concepts when a finite number of units are taken from a larger stock of material. As the number of units, n, increases, and $\bar{x} \to \mu$, $s \to \sigma$, the value of k'_2 decreases and eventually settles down to 2,0 (actually 1,96) when $p = 0{,}95$. Over the years, however, some measurement scientists have resisted the use of this concept. The more conservative ones use a factor of 3,0 and sometimes the range between maximum and minimum is used.

6 General principles of certification

There are a number of technically valid approaches to certifying a reference material. Valid approaches include measurement by one or more methods involving one or many laboratories. Depending on the type of reference material, its end-use requirements, the qualifications of the laboratories involved, and the quality of the method or methods, one approach may be chosen as more appropriate. Two of the more important aspects of the certification process involve the concepts of accuracy and uncertainty of the values determined for the properties being certified.

6.1 Certification on the basis of accuracy

Where technically possible, RMs are generally certified on the basis of accuracy.[23] Thus, a certified value generally represents the present best estimate of the "true" value. In some cases, the measurement cannot be in terms of a true value, so that an assigned property value for use with a specified method is adopted. The certification of the RM then requires no measurement campaign[1], but merely a statement of the assigned value and of the relevant measurement technique for which the CRM is a calibrant.

Certified values are not expected to deviate from the "true" value by more than the stated measurement uncertainty. The stated uncertainty of the value of a property must take into account any systematic and random errors inherent in the measurement process as well as any material variability and must describe them inclusively or separately. A number of different measurement approaches are commonly used by certifiers. They include the following.

a) Measurement by a single definitive method, (as defined in clause 7) in a single laboratory. The method is usually performed by two or more analysts working independently. Frequently, an accurately characterized back-up method is employed to provide additional assurance that the data are correct.

b) Measurement by two or more independent reference methods in one laboratory. The methods must have small estimated inaccuracies relative to the end-use certification requirement.

c) Measurement by a network of qualified laboratories using one or more methods of demonstrated accuracy.

In many cases various combinations of these approaches are used in certification. There are numerous advantages and disadvantages to each procedure for certification. The most important consideration of all of these approaches is that the systematic errors inherent in the methods used shall be well characterized and, to the extent possible, minimized. Systematic errors, precision of the test method, material variability and material stability must all be understood and taken into account when deriving the uncertainty statement for a certified property of an RM.

6.2 Certified values and their meanings — Uncertainties

In the development of each CRM, assurances must be obtained that the material used is uniform and stable, that test methods yield repeatable and consistent results, and that the conditions under which the material is to be used are carefully described. Eventually, these qualitative statements will have to be translated into quantitative terms, using data generated from the tests, and condensed into a certificate that will be understandable and useful to the user.

This condensation of information is no easy task! Ideally, the experimental conditions could be described in detail and all of the numerical values of individual determinations could be presented so that the user could judge how best to use these results.

Generally, the cost and work involved in a detailed presentation is not justified in relation to the number of times it will be profitably used. In most cases, therefore, the data are processed and condensed into the form presented on the certificate. The numerical values are normally expressed in two parts: the certified value of the property and the uncertainty of this value.

1) It is assumed here that the "specified method" is easily transferable. This is not true in some cases.

The uncertainty of the certified value denotes how well this value is known. A number of different expressions have been used for the statement of uncertainty, depending on how the CRM was developed, the group of scientists involved, and the use for which it was intended.

In many cases, the statements of uncertainty are based, to a certain extent, on the subjective judgement of the scientists involved, rather than on a strict interpretation of the data. Acquisition of data is expensive, sometimes prohibitively so, and these factors must be weighed against the intended use of the CRM.

ISO Guide 31 [3] is intended to assist CRM producers to prepare clear and concise certificates. The certificate should communicate essential information about a reference material from the producer to the user; in essence, this information is a statement of the certified property values, their meaning, and their limits of uncertainty. The remainder of the information is peripheral to this central statement and has two purposes: to describe the general nature and use of the material, and to assure the user of its integrity.

ISO Guide 31 recommends that the following information be contained in a certificate:

1) name and address of the certifying organization;
2) title of the document;
3) status of the certificate;
4) name of material;
5) sample number and/or batch number;
6) date of certification;
7) availability of other forms/sizes of the reference materials;
8) source of the reference material;
9) supplier of the reference material;
10) preparer of the reference material;
11) description of the reference material;
12) statement of intended use;
13) stability, transportation and storage instructions;
14) special instructions for correct use;
15) method of preparation;
16) statement of homogeneity;
17) certified property values and their uncertainty;
18) secondary property values given for information but not certified;
19) special values obtained by individual laboratories or methods;
20) meaning of the statistical uncertainty;
21) measurement techniques used for certification;
22) names of analysts, investigators, and participating laboratories;
23) legal notice;
24) reference (including companion report if any);
25) signatures or names of certifying officers.

More information and discussion of these points are contained in ISO Guide 31. Other useful references on certification of reference materials include NBS Special Publication 408 [24] and the Proceedings of the International Symposium on the Production and Use of Reference Materials [25].

7 Certification by a definitive method

7.1 Concept of definitive method as applied to reference materials

The certification of an RM by one measurement method requires the method to have high scientific status and a laboratory or laboratories of the highest quality. The method must be sufficiently accurate to stand alone for the determination of the property of interest. The actual accuracy of this method should be validated by international intercomparisons wherever possible. Such a method will have a valid, well-described theoretical foundation so that the reported results have negligible systematic errors relative to end-use requirements. The property in question is either directly measured in terms of base units of measurement or indirectly related to the base units through physical or chemical theory expressed in exact mathematical equations. If possible, the certifying laboratory should assure that the base units of measurements are traceable to appropriate national and/or international standards.

For the purpose of this Guide, a measurement method having these properties will be called a "definitive method" [16]. This definition of the term "definitive" is somewhat different from the definition given in the *International Vocabulary of Basic and General Terms in Metrology* [5]. However, the definition suggested here applies more directly to the certification of reference materials. Definitive methods are generally not practical for field work because they frequently require specialized equipment, they are often time consuming and expensive to perform, and they usually require highly skilled personnel. Thus, the acceptance of an RM certified in this way depends on the user community's confidence in the ability of the certifying laboratories to carry out the definitive method. Normal practice when a single certifying laboratory or organization is used is to require that the definitive method be performed by two or more analysts working independently, preferably using different experimental facilities.

The goal for writing an uncertainty statement is to present a clear, concise and objective summary of what is known about the material as a result of the testing programme. The exact form of the uncertainty statement will depend on the nature of the material, the needs of the user community for the particular CRM, and on the results of the testing programme. The outcome of the homogeneity assessment is especially important in determining the way uncertainties are expressed. Two important cases to be considered are

a) the material inhomogeneity is negligibly small compared to the measurement error (see 7.2);

b) the inhomogeneity is the major source of uncertainty compared to the measurement error (see 7.3).

7.2 Confidence interval (homogeneous material)

7.2.1 Concept

When the results of the homogeneity assessment indicate that material inhomogeneity is negligible, all units of the CRM can be treated as having identical values for the certified property. Thus, the value for each unit is the same as the mean value for all units, and one needs to summarize the uncertainty in the estimate of that mean. The only source of uncertainty is the random measurement error, and an appropriate and widely used method for expressing this uncertainty is a confidence interval for the mean.

7.2.2 Statistical model

Suppose that the experimental data consist of n independent measurements of the property to be certified. We represent these as $X_1, X_2, \ldots, X_n$. The mathematical model is

$$X_i = \mu + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

where

μ is the true mean value of the property measured;

ε_i is the measurement error associated with the ith measurement.

In this analysis, the ε_i are taken to be independent normal random variables with zero means and common (unknown) variance.

7.2.3 Transformation of data

In some cases it may be necessary to transform the data (for example by logarithms) in order to satisfy the assumption of normally distributed errors. In that case μ must be interpreted as the mean of the transformed value of the property of interest, and the final summary of the data involves retransforming the resulting confidence interval back to the original scale of measurement. In other cases where lack of a normal distribution is a problem, robust or non-parametric statistical procedures may be used to obtain a valid confidence interval for the quantity of interest.[26]

7.2.4 Description

A 95 % confidence interval has the form

$$\bar{x} \pm ts/\sqrt{n}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$$

$t = t_{0,975}\,(n-1)$ = the 0,975 fractile of the Student distribution with $(n-1)$ degrees of freedom.

To obtain a confidence interval for a confidence level other than 95 %, a different value of t is used. For example a 99 % confidence interval is obtained by taking $t = t_{0,995}$ for the appropriate degrees of freedom. It will be observed that for any given degree of freedom, $t_{0,995}$ is a larger number than $t_{0,975}$. Thus, for a given set of data, a 99 % confidence interval will always be wider than one computed at the 95 % confidence level. This is in accord with the intuitive notion that enlarging a given interval around $\bar{x}$ will increase the likelihood that the true mean μ will be contained somewhere in the interval. Detailed discussion of confidence intervals may be found in ISO 2602[1], including tables of values of t and $t/\sqrt{n}$.

7.2.5 Example of confidence interval

A highly pure arsenic trioxide (As_2O_3) material was certified as a CRM by a high accuracy coulometric assay method. Experimental work consisting of $n = 23$ determinations of mass fraction yielded $\bar{x} = 0{,}999\,893$ g/g and $s = 0{,}000\,104$ g/g. The data from this experiment vary symmetrically about the mean value, and include some individual values in excess of 1,00 g/g. The actual measurement technique has the property that valid measurement results can lead to calculated values for purity in excess of 100 % even though the actual purity must, of course, be less than 100 %. Thus, the nature of the data does not contradict the use of a statistical confidence interval based on the normal distribution.

The experiment involved duplicate analyses of 10 samples of the material and a triplicate of an 11th sample. The within- and between-sample variabilities were of equal magnitude, confirming that the material was homogeneous. ISO 2602 gives the value $t_{0,975}/\sqrt{n} = 0{,}432$ for $n = 23$. Therefore, a 95 % confidence interval for the (coulometric) purity of this material is

$$\bar{x} \pm (t/\sqrt{n})s = 0{,}999\,893 \pm (0{,}432) \times (0{,}000\,104)$$

$$= (0{,}999\,893 \pm 0{,}000\,045)\ \text{g/g}$$

The user of this material can understand that the purity of the material is almost surely inside these limits; that is, the mass fraction is between 0,999 848 and 0,999 938. Speaking more precisely, if this entire experiment were performed a large number of times (each with $n = 23$) and if 95 % confidence intervals were computed from the results of each experiment, then only 5 % of the intervals generated would fail to cover the true purity. This interpretation assumes that the systematic error of the measurement process is negligible.

1) ISO 2602, *Statistical interpretation of test results — Estimation of the mean — Confidence interval.*

7.3 Statistical tolerance interval (inhomogeneous material)

7.3.1 Concept

For some materials, it may be impossible or impractical to control the unit-to-unit variation to a level which is negligible compared to measurement uncertainty. In fact, the reverse may be true : measurement uncertainty may be negligible compared to between-units variation. In this situation the individual units of the CRM constitute a population for which the property of interest varies slightly from one unit to the next. The uncertainty statement on the CRM certificate must then describe the extent of variation among units in the population so that a user can be confident that any unit he might obtain will have a value somewhere in the stated range. The issuer of a CRM must determine whether the inherent between-units variability is sufficiently small that the material will be useful for its intended purpose, for, if not, there is no reason to proceed with certification.

7.3.2 Computation

In order to generate a statistical tolerance interval which properly reflects the unit-to-unit variability of the material, the data $X_1, X_2, \ldots, X_n$ should consist of n independent values — one for each of n **distinct** units. The mathematical model is

$$X_i = \mu + \beta_i \qquad i = 1, 2, \ldots, n$$

where

μ is the mean value of the property, averaged over the population of all units;

β_i is the difference between the value of the property for the ith unit and the population mean.

Typically the quantity β_i will also reflect some unavoidable contribution of the random measurement error realized in measuring the ith unit. The β_i are taken to be independent normal random variables with zero means and common (unknown) variance. As discussed in 7.2.3, it is assumed that the raw data have been transformed, if necessary, to achieve this condition on the β_i. The computation of a tolerance interval was described in 5.5.

8 Certification by interlaboratory testing

8.1 General concept and practice

The concept of the certification of an RM by interlaboratory test is based on at least two assumptions :

a) there exists a population of laboratories that is equally capable in determining the characteristics of the RM to provide results with acceptable accuracy;

b) assumption a) implies that the differences between individual results, both within and between laboratories, are statistical in nature regardless of the causes (i.e. variation in measurement procedures, personnel, equipment, etc.).

Each laboratory mean is considered to be an unbiased estimate of the characteristic of the material. Usually, the mean of laboratory means is assumed to be the best estimate of that characteristic; however, in the case of very irregular distributions such as may be found for example in trace element analysis, the use of a more robust statistic such as the median or a trimmed mean may be appropriate.

In practice, the size of the laboratory population that is available to an interlaboratory analysis programme is limited. In most cases, therefore, a random-design model cannot be fully implemented.

8.1.1 General procedure

The general procedure for the certification of an RM by interlaboratory consensus is outlined schematically in figure 3. Each stage can be treated as being distinct and possesses criteria that must be satisfied before proceeding to the next stage.

8.1.2 Confirmation of homogeneity as part of interlaboratory programme

The results of the interlaboratory programme can serve as a final confirmation of the homogeneity of an RM provided that a two-way nested design has been followed in which pq units are used and p laboratories each determine the value of the characteristic of q units with n replicate determinations per unit. It is important that the laboratories do not deviate from the design.

8.2 Organization of interlaboratory programme

To be successful, an interlaboratory programme must have a well-defined objective, be effectively designed and be efficiently organized with clear, concise guidelines with which participating laboratories can readily comply. Participation in such a programme implies an agreement to adhere to these guidelines. These guidelines consist of time objective, number of units, number of replicate determinations per unit, measurement methods, test portion size where applicable, etc.

8.2.1 Time objective

The organizer must set the time schedule, i.e. the dates when the samples are to be distributed and when the test results are to be reported.

8.2.2 Number of participating laboratories

The minimum number of participating laboratories comprising an interlaboratory programme for the characterization of an RM varies with the complexity of the necessary measurement procedure. The more complex the procedure, the larger is the between-laboratories variation to be expected, thereby necessitating an increasing number of participating laboratories to achieve a consensus value having a predetermined precision. In practice unfortunately, the more complex the procedure, the fewer are the laboratories capable of performing the procedure. In extreme cases, the certifying agency may be forced to forego an interlaboratory programme altogether for certain specialized candidate RMs.

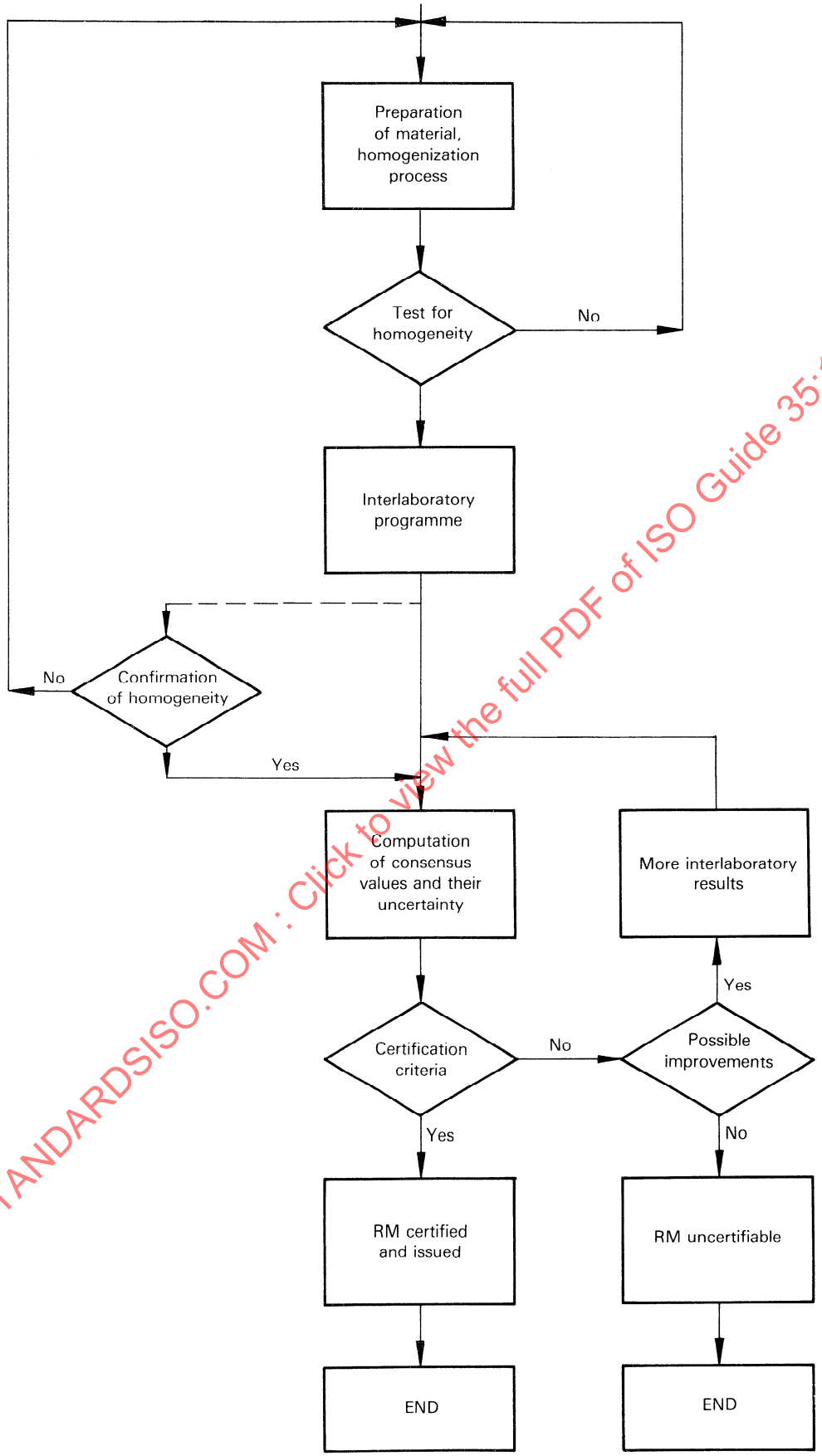


Figure 3 — Schematic diagram of the process of preparation and certification of an RM by interlaboratory consensus

There is general agreement among laboratories experienced in interlaboratory testing that the preferable number of participating laboratories is 15 or more.

8.2.3 Number of units and replicate determinations

If the results of the interlaboratory programme are to serve as a final confirmation of the homogeneity of an RM, values of the characteristic for a minimum of two units of the RM should be determined by each participating laboratory. Otherwise, one unit for each participating laboratory may be sufficient.

The minimum number of replicate determinations is two per unit of RM. All replicate determinations should be made on separate test portions.

8.2.4 Measurement methods

The organizer of an interlaboratory programme may specify the use of a single method to participating laboratories when such a well-established "standard" measurement procedure is available. Otherwise, the organizer should allow each participating laboratory to use the method of its choice, provided that he has evidence of the validity of such methods.

8.2.5 Reporting results

Participating laboratories should report individual results (not the average). The number of significant figures reported should comply with the guidelines for the programme. It is recommended that an outline of the measurement procedure used be reported in sufficient detail to permit an understanding of all preliminary stages in the measurement process, for example in chemical analysis the decomposition of the sample and separation of the analyte(s) of interest. Reference to the literature where appropriate should be stated.

8.3 Initial processing of the results

The results submitted by the participating laboratories are evaluated in accordance with the procedure outlined in figure 4.

8.3.1 Presentation of the results

For the convenience of processing and for future reference, the results from an interlaboratory programme should be grouped on the basis of the characteristic and tabulated systematically. This table should include identification of the laboratory and the method, individual results, laboratory mean and corresponding standard deviation. However, if the participating laboratories determined the value of the characteristic for more than one unit of RM, it is recommended that the within-unit and overall mean and corresponding standard deviations be presented in a table separate from the individual results. When a participating laboratory has submitted more than one set of results for a characteristic obtained by different measurement methods, each set should be treated independently, i.e. as if from another laboratory.

It is also recommended that the results be presented in graphical form.

8.3.2 Technically explainable outliers

The results must be checked for technically explainable outliers which are excluded before any statistical evaluation is performed. If possible, it is recommended that the participating laboratory concerned be informed for its benefit and be invited to submit new results as replacements.

8.3.3 Minimum number of laboratories

After the exclusion of outliers, the remaining number of sets of results must be consistent with the principle of the minimum number of laboratories necessary to comprise an interlaboratory programme.

8.3.4 Frequency distribution of results

It is essential to know how the results are distributed. In many cases, the distribution can be observed graphically.

8.3.4.1 Multimodal distribution

If most of the results form two or more clusters, no consensus value can be inferred. The following possibilities should be considered:

a) if there is correlation of these clusters with measurement method procedures, and if the difference between the means of these clusters is both statistically and physically significant, then there is no consensus value; in this case, improvement in the measurement method procedures is necessary to resolve the problem;

b) if there is no correlation of these clusters with measurement method procedures, and if the difference between these clusters is statistically and physically significant, a larger pool of results may be necessary to overcome the relatively poor measurement methods available.

8.3.4.2 Unimodal distribution

If most of the results form a single cluster, it can be inferred that a consensus value does exist.

If the distribution is unimodal, a decision should be made as to whether an assumption of normality is reasonable. This decision can be based either on a visual observation of the histogram, on the normality test or on past experience with the nature of the determinations.

As discussed in 7.2.3, in some cases the results have to be transformed into another form where they can be assumed to follow a normal frequency distribution. Some commonly used transformations include logarithmic, square root and exponential forms.

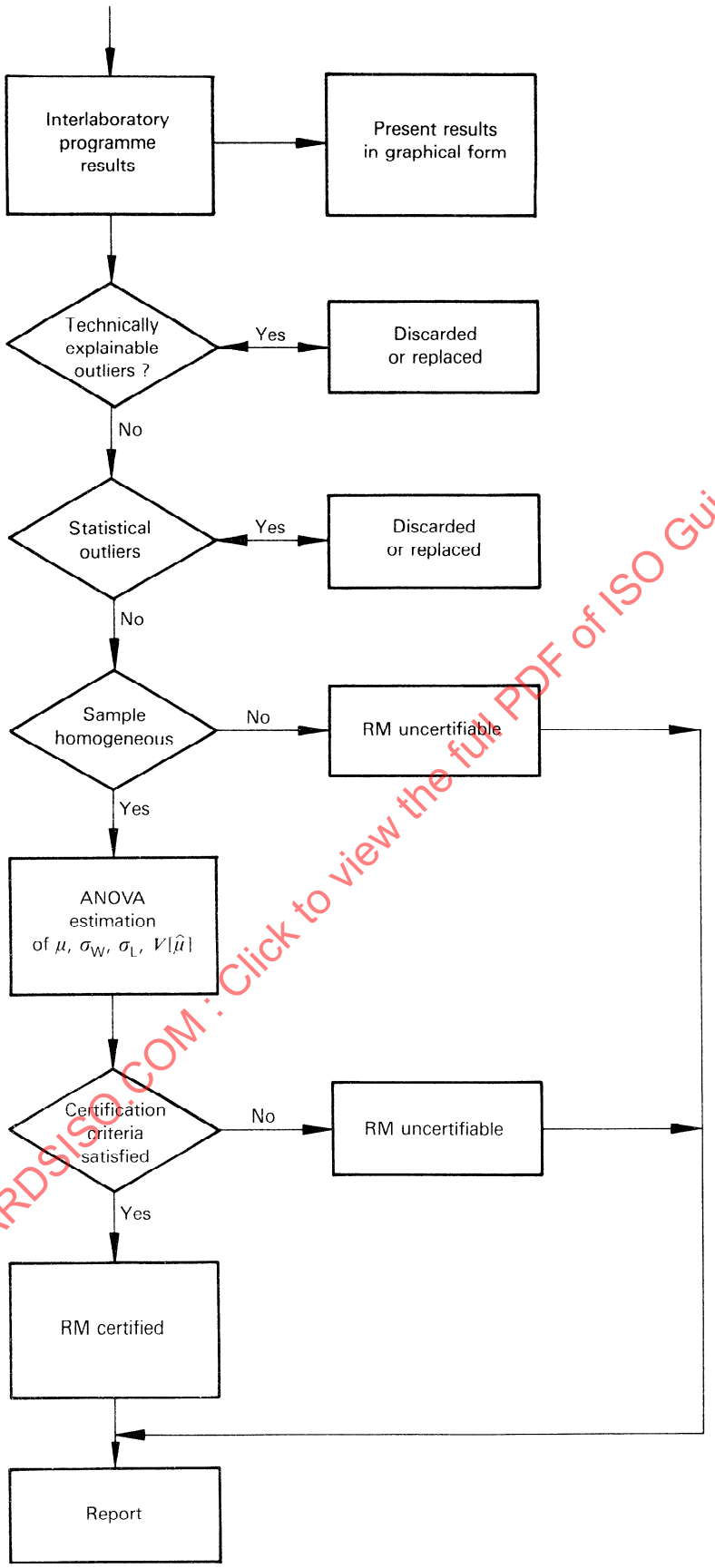


Figure 4 — Schematic diagram for statistical evaluation of interlaboratory results for certification of an RM

8.3.5 Statistical outliers

A single result or an entire set of results is suspected to be a statistical outlier if its deviation either in accuracy or precision from others in the set or other sets, respectively, is greater than can be justified by statistical fluctuations pertinent to a given frequency distribution. Therefore, the effectiveness for the detection of outliers depends on the validity of the assumption of the frequency distribution. The test for outliers should be the statistician's prerogative. For an interlaboratory programme outlying status may be conferred on individual results, results for individual units or the entire set of results from a laboratory.

8.4 Statistical analysis

8.4.1 Two-stage nested design

This model is used when the results of an interlaboratory programme are used to confirm the homogeneity as well as to characterize the material. The experimental scheme is illustrated schematically in figure 5 a). The results can be expressed by the equation

$$X_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk} \qquad \ldots (5)$$

where

X_{ijk} is the kth result of sample unit j reported by laboratory i;

μ is the grand mean;

α_i is the error due to laboratory i;

β_{ij} is the error due to the jth sample unit in laboratory i;

ε_{ijk} is the measurement error.

8.4.2 One-stage nested design

This model is used when the material is accepted to be homogeneous by the organizers. The experimental scheme is illustrated schematically in figure 5 b). Equation (5) can then be simplified to

$$X_{ik} = \mu + \alpha_i + \varepsilon_{ik}$$

8.4.3 Analysis of two-stage nested design

Parameters to be estimated are

— μ, the grand mean (which is used as the consensus value);

— σ_L^2, the variance of the between-laboratories error (α_i);

— σ_U^2, the variance due to between-units inhomogeneity (β_{ij});

— σ_W^2, the variance of the within-laboratory measurement error (ε_{ijk}).

All these parameters can be estimated simultaneously by the analysis of variance (ANOVA) method (see 8.4.3.1) if there are sufficient results of equal replication (the same number of replicate determinations from each unit and the same number of units per laboratory) after outliers have been excluded. If this ANOVA requirement cannot be met because of the number of outliers and/or missing results, the significance of the between-units (inhomogeneity) variance can be tested by the simple procedure for unbalanced data given in 8.4.3.2.

Theoretical details and additional methods for balanced and unbalanced ANOVA are given in standard textbooks. [27, 28]

8.4.3.1 Computation of two-stage ANOVA

x_{ijk} is the kth result of sample unit j reported by laboratory i;

p is the number of participating laboratories;

q is the number of units per laboratory;

n is the number of replicate determinations per sample unit.

$$\bar{x}_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk}$$

$$\bar{x}_i = \frac{1}{q}\sum_{j=1}^{q} \bar{x}_{ij}$$

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

The sums of the squares SS_1, SS_2 and SS_3 are calculated by the following equations :

$$SS_1 = qn\sum_{i=1}^{p} (\bar{x}_i - \bar{x})^2$$

$$SS_2 = n\sum_{i=1}^{p}\sum_{j=1}^{q} (\bar{x}_{ij} - \bar{x}_i)^2$$

$$SS_3 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2$$

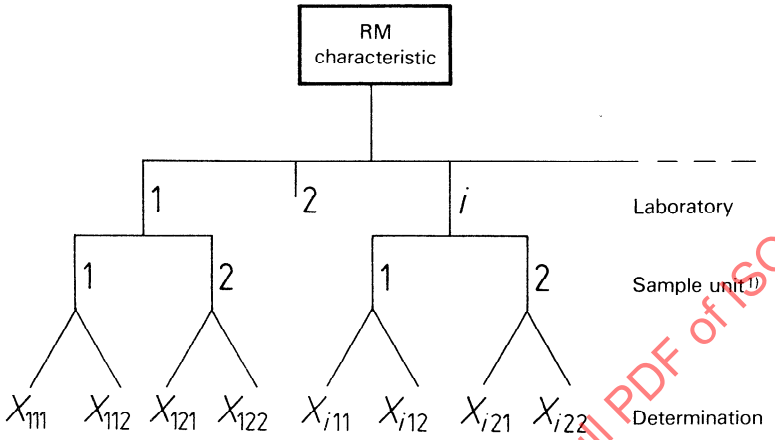
The degrees of freedom are

$$f_1 = p - 1$$

$$f_2 = p(q - 1)$$

$$f_3 = pq(n - 1)$$

a) **Two-stage nested design**



1) All sample units are different. However, in each laboratory they are numbered 1, 2, ...

b) **One-stage nested design**

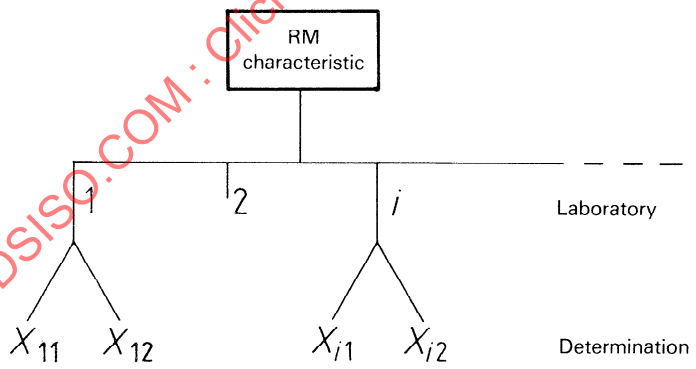


Figure 5 — Experimental scheme for an interlaboratory programme

and each mean square is given as

$$MS_1 = SS_1/f_1$$
$$MS_2 = SS_2/f_2$$
$$MS_3 = SS_3/f_3$$

These results should be tabulated (see table 1).

Table 1 — ANOVA table

Source	Sum of squares	Degrees of freedom	Mean square	Expectation of mean square
Between laboratories	SS_1	$p - 1$	MS_1	$\sigma_W^2 + n\sigma_U^2 + qn\sigma_L^2$
Between units	SS_2	$p(q - 1)$	MS_2	$\sigma_W^2 + n\sigma_U^2$
Measurement error	SS_3	$pq(n - 1)$	MS_3	σ_W^2

Each parameter is estimated by the following equations, where the circumflex denotes the estimate :

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma}_L^2 = (MS_1 - MS_2)/qn$$
$$\hat{\sigma}_U^2 = (MS_2 - MS_3)/n$$
$$\hat{\sigma}_W^2 = MS_3$$

If the numerical value of $\hat{\sigma}_L^2$ or $\hat{\sigma}_U^2$ is negative, zero should be used instead.

The tests for statistical significance are

a) between-units (inhomogeneity) variance

$$F_{2|3} = MS_2/MS_3$$

which should be compared with the critical value of the F-distribution for degrees of freedom $p(q - 1)$ and $pq(n - 1)$;

b) between-laboratories variance

$$F_{1|2} = MS_1/MS_2$$

which should be compared with the critical value of the F-distribution for degrees of freedom $(p - 1)$ and $p(q - 1)$.

The variance of the consensus value $\bar{x}$ is estimated by

$$\hat{V}(\bar{x}) = \frac{MS_1}{pqn}$$

The confidence interval for μ based on $\bar{x}$ is from A to B where

$$A = \bar{x} - t_{1-\alpha/2}(p - 1)\sqrt{\frac{MS_1}{pqn}}$$

$$B = \bar{x} + t_{1-\alpha/2}(p - 1)\sqrt{\frac{MS_1}{pqn}}$$

where $t_{1-\alpha/2}(p - 1)$ is the $1 - \alpha/2$ fractile of the t-distribution with $(p - 1)$ degrees of freedom.

8.4.3.2 Modified ANOVA for unbalanced data

x_{ijk} is the kth result of sample unit j reported by laboratory i;

p is the number of participating laboratories

q_i is the number of units at laboratory i;

n_{ij} is the number of replicate determinations of sample unit ij.

$$\bar{x}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk}$$

$$\bar{x}_i = \frac{\sum_{j=1}^{q_i} n_{ij}\bar{x}_{ij}}{\sum_{j=1}^{q_i} n_{ij}}$$

The sums of the squares SS_2 and SS_3 are calculated by the following equations :

$$SS_2 = \sum_{i=1}^{p}\sum_{j=1}^{q_i} n_{ij}(\bar{x}_{ij} - \bar{x}_i)^2$$

$$SS_3 = \sum_{i=1}^{p}\sum_{j=1}^{q_i}\sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij})^2$$

The degrees of freedom are

$$f_2 = \sum_{i=1}^{p} (q_i - 1)$$

$$f_3 = \sum_{i=1}^{p}\sum_{j=1}^{q_i} (n_{ij} - 1)$$

and the mean squares are given as

$$MS_2 = SS_2/f_2$$
$$MS_3 = SS_3/f_3$$

These results should be tabulated (see table 2).

Table 2 — ANOVA table

Source	Sum of squares	Degrees of freedom	Mean square
Between units	SS_2	f_2	MS_2
Measurement error	SS_3	f_3	MS_3

The test for statistical significance of the between-units (inhomogeneity) variance is

$$F_{2|3} = MS_2/MS_3$$

which should be compared with the critical value of the F-distribution for degrees of freedom $\left\{\sum_i (q_i - 1)\right\}$ and $\left\{\sum_i \sum_j (n_{ij} - 1)\right\}$.

8.4.4 Analysis of one-stage nested design

For cases where the material is considered to be homogeneous, i.e. that all units are identical, all results reported by a laboratory are considered as replicates.

x_{ij} is the jth result reported by laboratory i;

p is the number of participating laboratories;

n_i is the number of results reported by laboratory i.

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ijk}$$

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

The variance of the consensus value, $\bar{\bar{x}}$ is simply estimated by

$$\hat{V}(\bar{\bar{x}}) = \frac{1}{p(p-1)} \sum_{i=1}^{p} (x_i - \bar{\bar{x}})^2$$

with degrees of freedom $(p - 1)$.

The confidence interval for the consensus value (mean of means) is the interval from A to B where

$$A = \bar{\bar{x}} - t_{1-\alpha/2}(p - 1) \left(\hat{V}(\bar{\bar{x}})\right)^{1/2}$$

$$B = \bar{\bar{x}} + t_{1-\alpha/2}(p - 1) \left(\hat{V}(\bar{\bar{x}})\right)^{1/2}$$

and $t_{1-\alpha/2}(p - 1)$ is as described in 8.4.3.1.

9 Certification based on a metrological approach

9.1 Concepts

The objective of this approach is to produce certified values the accuracy and the uncertainty of which are demonstrated by experimental evidence.

The first basic concept behind this approach is that when the property, physical or chemical, of a material can be defined from first principles, its value does not depend on a particular method used for the measurement.[1)] When the value of such a property is to be certified, it is therefore important for the certification body to show that the value does not include a systematic error specific to a method or to a laboratory. The procedure consists in measuring the property under consideration by different methods which are considered to be the most accurate in the actual state of the art and applied by laboratories most experienced for the respective methods. This approach is also adopted by establishments working alone : they use several methods, possibly with operators working independently, and compare the results.

The second concept is that the uncertainty statement, which is an important part of the value assigned to a measurement standard, can fail to be reliable when it is not based on a very careful comparison between results of different (high level) laboratories and different methods. This is illustrated by examples in 9.2 and 9.3.

The measurement of the quantities referred to above is traceable or should be traceable to measurement scales, themselves traceable to the SI. By definition, the traceability is the property of a result of a measurement whereby it can be related to appropriate standards through an unbroken chain of comparisons.

The traceability is necessary to support the concept of accuracy. The traceability of analytical processes is more difficult to establish than in physical measurements. The problems involved in this traceability are discussed in detail in 9.3.

In 9.4, examples are given of properties which are defined only by a method and can be traceable only to a conventional measurement scale.

9.2 Certification of physical properties

The most accurate measurements are carried out for fundamental units, their most common multiples and their submultiples, in the primary metrology laboratories. Here, all sources of errors and uncertainties are investigated in great detail; methods of measurement have been improved over many years to reduce uncertainties. The accuracy of these measurements is fairly well established, especially when they have been the subject of interlaboratory comparisons. Reservations must be made for measurements where there has been no intercomparison. In addition, any new laboratory being established needs extensive intercomparisons to ensure that its

1) There are properties which are defined only as a function of a method; this question is examined in 9.4.

own estimates of accuracy are correct and that no error has escaped its attention.

Intercomparisons add confidence to the uncertainty computed by the metrology laboratories individually. Sometimes they use safety factors which are not necessary; sometimes they underestimate their own uncertainties.

The present practice by which each metrology laboratory evaluates the uncertainty of a particular measurement on its own is inherently dangerous. It is not possible for a laboratory alone to avoid all errors in all circumstances, in particular for derived units. Intercomparisons detect errors that were not taken into account and situations where all parameters influencing the measurements are not sufficiently well controlled.

There is unfortunately no general requirement in metrology that uncertainty statements be based on appropriate intercomparisons. Certifying a reference material on the basis of results of one single metrology laboratory may therefore imply a risk which should not be overlooked.

When the certification of a physical property or quantity is undertaken, it is therefore important to have an intercomparison between the major metrology laboratories followed by a full discussion of the results with all participants to resolve any possible discrepancy. If the primary metrology laboratories are not themselves involved in the measurement, complete traceability of the participating laboratories to the respective national laboratories must be established before starting.

The participants must then compare their measurements and discuss all the possible errors responsible for discrepancies and eliminate them while remaining independent. This is described in more detail for chemical measurements in 9.3.2.

If more than one method is possible, and if these methods appear equally valid, it is important to compare them. However, it is useful to remember that the method with the shortest traceability route or, in other words, with the most direct connection to the fundamental units, has a higher probability of being more accurate.

At the limit, there can of course exist situations where one single laboratory, having compared its method with all possible others and having eliminated most causes of errors, is able to refine its method to reduce the uncertainty while taking considerable precautions to avoid any accidental source of errors.

Some measurement problems in the field of physical properties can be briefly illustrated by thermal conductivity of insulation and refractory materials. Until some years ago, laboratories were not able to carry out such measurements with appropriate accuracy although the calibration of the instrumentation appeared satisfactory. The guarded hot-plate used for the measurement was constructed and operated in accordance with existing national and international standards. The agreement appeared satisfactory for simple technical applications. However, in most laboratories there was a systematic error. Heat losses occurred above room temperature because the guard ring was not sufficient. Any reference material certified on that basis would have a wrong traceability. The method and equipment were therefore modified until the heat losses became negligible.

The accurate determination of thermal conductivity of refractory materials is very difficult by the direct method using the guarded hot-plate apparatus mainly because of the heat losses and experimental difficulties. Methods such as the hot-wire method or the flash method do not present such difficulties, but their traceability is not easy to establish and therefore these methods are not the best for certification. However, the results of these methods are important as a verification of the results of the guarded hot-plate.

9.3 Certification of a chemical composition

9.3.1 Traceability

In the field of analytical chemistry, there is no established measurement system organized as in the field of metrology, with primary and calibration laboratories, and measurement standards available for circulation. The concept of accuracy is hence more difficult to reach and the traceability is more difficult to realize.

In chemistry, the calibrations in the usual sense are not the major source of difficulties although the task of the chemist is heavier than that of the metrologist. He needs not only physical standards of mass, volume, temperature, etc. but also standards of all chemical species he has to determine : elements, organic compounds, etc. Each one of these chemical standards has an uncertainty (e.g. impurities) which is sometimes underestimated.

The biggest problem is however the traceability of the overall analytical process : the traceability chain is broken every time the sample is physically or chemically modified in the analytical process.

As the variety of sample processing procedures is large, it is not possible to discuss the traceability in general. The following paragraphs are to be considered only as examples.

9.3.1.1 Sample weighing

The first step of the analytical process is the weighing of the sample. This does not pose problems of traceability if the balance is periodically calibrated. Human errors are not excluded but they are not frequent.

9.3.1.2 Sample treatment

Whenever the sample is dissolved or submitted to similar treatment, the traceability chain is broken and any uncertainty evaluation should take this into account. To establish traceability for that part of the measurement procedure, a laboratory must demonstrate the relationship between the initial sample and the solution prepared from it. The main questions to be answered are, was the sample totally dissolved, what were the losses, were there contaminations? If the analysis is to determine not one element but a compound, was the compound changed during the dissolution step? In the case of organic compounds, the efficiency of extraction is one of the main causes of difficulties.

If, after sample treatment, the solution is subject to further manipulations (preconcentration, precipitation, etc.) each step complicates the traceability route and adds new possibilities of losses or contaminations which must be investigated.

It is well known that some of the parameters listed here depend more on the matrix than on the element or compound to be determined.

9.3.1.3 Final determination

The third step in an analytical process is the final determination. Apart from gravimetry, titrimetry, and coulometry, most methods, for example spectrometry and atomic absorption, are indirect. The instrumentation used for these measurements provides a signal which must be correlated with the concentration of the substance of interest in the unknown sample. That correlation is established by means of a calibration curve.

Here there are two groups of problems to consider :

- is any error introduced in producing the calibration curve and what is the accuracy?
- is it correct to use that particular calibration curve?

If we suppose that the calibration can be done by means of solutions, then the most important parameters to take into account are

- the accuracy of the measurements (mass, volume) made for the preparation of the solution;
- the purity of the elements or substances, the stoichiometry of the compounds, etc.;
- the purity of the water or solvent.

Errors due to the calibration curve are not rare even in good laboratories.

However, as pointed out in 9.3.1.4 even larger errors are due to the fact that users sometimes produce calibration curves which are not appropriate to the solutions they have to analyse; these are named matrix effects, interferences, etc.

In metrological terms, this could be expressed as follows : each laboratory produces for itself a measurement scale which is not fully appropriate to the measurements to be made, and each one produces a different measurement scale.

9.3.1.4 Matrix effect

The response of a particular element to a measurement process (e.g. spectrometry, atomic absorption) may depend on the solution (viscosity, conductivity, ionic strength) or on the ions present in it (interferences).

Besides a large number of such cases in inorganic analyses, severe matrix effects are found in clinical chemistry, where some methods designed to analyse a serum can be wrong for aqueous solutions. For such methods the calibration should be done with human serum; if this is not possible, the validity of any other matrix should be demonstrated.

In this respect the term "calibrant" used by biochemists can be misleading. Similarly, in inorganic chemistry, a calibration solution should simulate very closely the solution to be analysed.

9.3.2 Certification work

The task of any laboratory participating in an exercise to certify a new reference material includes the study of the parameters mentioned in 9.3.1. A full study requires the comparison of different methods of sample treatment and different methods of determination. This can, however, be best done collectively in order to have the collaboration of experienced specialists in each method. In addition, for each method there should be more than one laboratory in order to avoid systematic errors due to laboratory effects or operator effects. It can be pointed out that errors (e.g. those due to contaminations) can only be detected by comparison of results from different laboratories.

The need for scrutinizing carefully the results of the different participants can be illustrated by the examples given in tables 3 and 4, which are rather typical of trace element analysis at very low levels. The laboratories often find values which are too high because they all produce some contamination. If one too quickly adopted the mean value of their results, one would have a systematic error by excess, and a reference material totally unreliable from the point of view of traceability. This explains why the procedure proposed to approach accuracy is composed of several steps in which the participants discuss all sources of errors in all parts of the analytical procedure and then try to reduce them. Analyses are then repeated (possibly not on exactly the same samples) and the results are discussed again as many times as necessary to reach sufficient convergence.

The need for several laboratories also exists in the case of so-called "definitive" methods like IDMS. For one particular determination there may be more than one "definitive" method, or several variations of a definitive method; it is of course essential to verify that they provide the same result and this is not necessarily the case. If, after detailed comparison of the results of several laboratories, it is not possible to identify errors, the variation of results (between laboratories) represents the uncertainty of the technique in the current state of the art. Working with one single laboratory would perhaps lead to a smaller spread of results but this would not necessarily represent the real uncertainty.

To summarize, the certification work in accordance with the approach proposed here would include the following steps for a homogeneous and stable material :

- examination, with experienced laboratories, of the most reliable (accurate) methodologies for the analysis of the element or substance in the particular matrix considered;
- a first round of analyses;
- a detailed discussion of the results with all participants to try to discover explanations of the differences; particular attention is given to
  - sample treatment,
  - possible losses, contaminations,

Table 3 — Trace elements in milk

Values in nanograms per gram

Element	First intercomparison (range of results)	Certification campaign (range of results)	Certified
Cd	0,4 to 4 500	1 to 5,6	2,9
Hg	0,6 to 42	0,73 to 1,27	1,0
Pb	68 to 5 500	92,4 to 112,5	104,5
Cu	470 to 9 257	475 to 700	545

Table 4 — Results of analyses of olive-tree leaves

Element	1979 results µg/g	Ratio	1981 results µg/g	Ratio
Cd	0,050 to 6,654	133	0,054 3 to 0,121	2,2
Pb	17,6 to 33,3	1,9	20,2 to 26,4	1,3
Hg	0,005 to 0,702	140	0,247 to 0,336	1,4
Cu	0,5 to 131,9	264	43,2 to 50,8	1,2
Zn	12,3 to 31,6	2,6	14,5 to 17,7	1,2
Mn	0,4 to 4,6	11,5	51 to 61,8	1,2

Table 5 — Determination of pesticides in powdered milk spiked with certain compounds

Compound	Results mg/kg	Ratio	Quantities added mg/kg
HCH	0,001 to 0,22	220	0,28
α-HCH	0,009 to 0,60	67	0,11
γ-HCH	0,001 14 to 0,18	158	0,20
DDE	0,004 3 to 0,47	109	0,54
op'DDT	0,003 to 0,24	80	—
β-HCH	0,01 to 0,13	13	0,08
β-HEPO	0,001 to 0,13	130	0,12
Dieldrin	0,01 to 0,104	10	0,10
pp'DDT	0,005 to 0,36	72	—

— solution treatments,

— errors included in the calibration curve,

— matching the calibration to the product to analyse matrix effects, interferences;

— a second round of analyses with the same laboratories but possibly with a material of slightly different composition;

— discussion;

— further rounds of analyses as necessary.

The procedure described often leads to rejecting some method(s) or to abandoning some laboratories which cannot improve their performance. At the end of this long procedure, one has a set of technically consistent results for which one calculates the mean value, and its 95 % confidence interval (adopted as uncertainty). Examples of successive stages are given in figures 6 and 7. Statistics are used for no other purpose than for verifying that the conditions are fulfilled to calculate a 95 % confidence interval.

The statistics for the calculation are the same as shown in ISO Guide 33 [29].

When the results are not consistent, one must conclude that the technical work is not terminated and that certification is not possible.

It is to be noted that for trace elements or for the certification of impurity levels, the distribution of results can be log-normal. The confidence interval can be non-symmetrical.

NOTES

1 The method(s) used to certify a reference material are sometimes very different from the methods used in routine practice (e.g. to certify cortisol in serum one has to use GCMS, while in practice the commonly used method is radio-immunoassay). In these cases it is important to verify that the RM is suitable for use with the routine method.

In figure 9, it should be noted that only the GCMS results were intended for certification. The other methods were used to verify the suitability of the RM.

2 For the preparation of a reference material in the biomedical field in particular, blood serum is treated with stabilizing agents or is lyophilized. It is then essential to verify the appropriateness of the reference material after these treatments.

9.4 Certification of conventional properties

In chemistry, biochemistry and other technologies, many properties are defined only by a method, a test procedure or particular equipment. Examples are mechanical properties of materials, activity of enzymes, etc. The results of these measurements or tests can be subject to great variability with heavy economic consequences.

As in any other measurement, the results depend on the way in which the procedure is applied. However, the procedure is not always described in all necessary details in the written standards and the operator has no means of verifying if the way he has interpreted and applied the procedure is correct. Hence the need for the reference material.

The diagrams in figure 10 show results of determination of the activity of an enzyme (γ-glutamyltransferase) in an albumin matrix with the same IFCC method. Laboratories shown on the right-hand side had previous training with the method. Laboratories on the left-hand side were high-level scientific laboratories but with no previous experience in the method. While the two upper diagrams in figure 10 relate to one material, the bottom diagram concerns a different material.

Similarly, where a test depends on the use of a particular machine or equipment it is possible, but extremely time-consuming and expensive, to verify that the machine satisfies all specifications. A simple way to by-pass this is to measure or test a reference sample. If the results are satisfactory, it means that the machine is in good condition and that therefore the results can be considered traceable to the measurement scale established by the relevant written standard.

Of course, the certification work to establish reference materials for such properties or measurement scales requires the application of the same principles as explained before. The measurements of these parameters, which may be mass, volume, length or temperature, must themselves be accurate and traceable and therefore may require extensive calibration. Considerable effort is often necessary to investigate the influence of the various parameters of the procedures and of the equipment on the measurement results. The verifications and calibrations must be done independently in a few, if not several, laboratories in order to avoid a uniform bias that would appear as a good agreement and give an illusion of accuracy.

9.5 Use of reference materials for establishing traceability

In 9.3.1, a review was given of a number of parameters that a laboratory should control and verify to ensure the traceability of the determinations. To do this in all necessary details is very hard work.

This can be considerably simplified by the use of a certified reference material of established traceability. The reference material must be sufficiently similar (in matrix) to the actual sample to be analysed in order to include all analytical problems which might cause errors in the determinations. Of course, the user should apply to the reference material the same analytical procedure as for his unknown sample.

When the laboratory using such a reference material finds only a negligible difference with the certification value, this indicates both that the result is accurate and that it is traceable to the fundamental measurement scale. If the difference is not acceptable, it indicates that the measurement procedure includes errors which must be identified and eliminated. It is suggested that the most critical steps subject to errors are the sample treatment and the matching of the calibration.

Hence the role of the reference material is comparable to that of the transfer standards used in metrology laboratories in industry, in that it allows working with a specified margin of uncertainty.

The reference materials also make it possible to establish the uncertainty of a measurement for analytical determinations or technological testing.

The importance of a certified reference material goes therefore beyond the definition of the reference material given in ISO Guide 30 [2].

A reference material is used not only

- for calibration of an apparatus,
- for the verification of a measurement procedure,

but also

- for establishing traceability of the measurement results,
- for determining the uncertainty of these results.

Finally, one should not forget that the use of a reference material does not eliminate completely the importance of audits, the purpose of these being to verify that no mistake is made in the use of the RM.

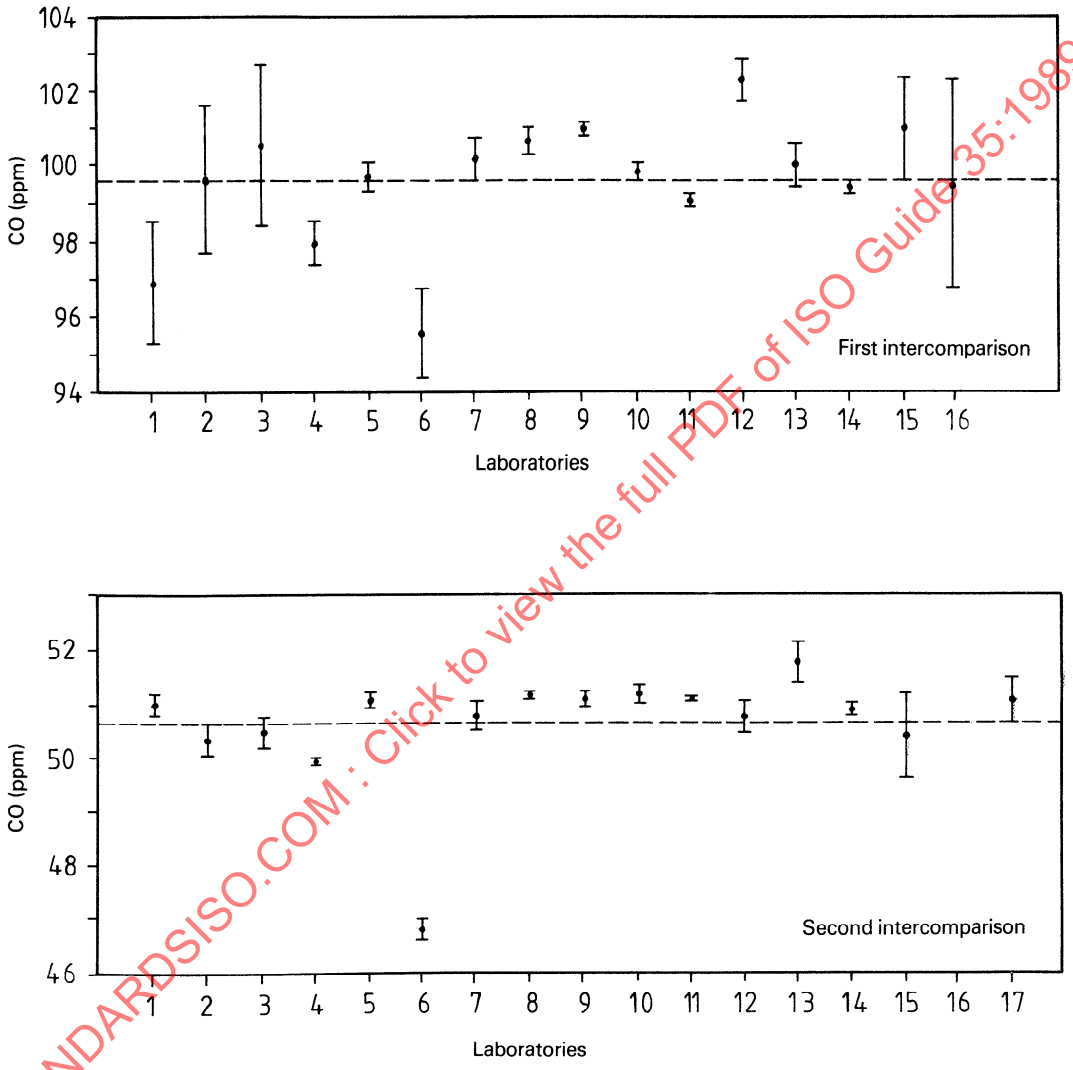


Figure 6 — Results of the first and second intercomparison of analyses of carbon monoxide in nitrogen