GUIDE 35

Reference materials — General and statistical principles for certification

Third edition 2006

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

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

Published in Switzerland

Contents

Page

Foreword

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

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

Draft Guides adopted by the responsible Committee or Group are circulated to the member bodies for voting. Publication as a Guide requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO Guide 35 was prepared by the ISO *Reference Materials Committee* (REMCO).

This third edition cancels and replaces the second edition (ISO Guide 35:1989), of which all clauses referring to the estimation of measurement uncertainty have been thoroughly revised. This revision also provides an up-to-date description of the technical issues related to the production and certification of reference materials.

Introduction

The production, characterization and certification of reference materials (RMs) is a key activity in improving and maintaining a worldwide coherent system of measurements. As detailed in ISO Guide 32 and ISO Guide 33, certified reference materials (CRMs) are used for calibration, quality control and method validation purposes, as well as for the assignment of values to other materials, which in turn can also be CRMs. Furthermore, CRMs are used to maintain or establish traceability to conventional scales, such as the octane number, hardness scales and pH. Last, but not least, selected pure substances are also used to maintain the international temperature scale.

For producers of CRMs, there are three ISO Guides that assist the set-up of a facility to produce and certify RMs and to ensure that the quality of thus-produced CRMs meet the requirements of the end-users. ISO Guide 34 outlines the requirements to be met by a CRM producer to demonstrate competence, whereas this Guide provides assistance on how to meet these requirements. At a fairly generic level, this Guide provides models for homogeneity testing, stability testing, and the characterization of the candidate CRM. ISO Guide 31 describes the format and contents of certificates for CRMs.

In some ways, this Guide can be seen as an application of the *Guide to the Expression of Uncertainty in Measurement* (GUM) with respect to the peculiarities of the production of CRMs. Where possible, this Guide makes reference to the GUM, as the latter describes in detail how to evaluate measurement uncertainty of a value obtained from measurement. This Guide complements the GUM in a sense that it provides additional guidance with respect to the inclusion of the uncertainties due to the (remaining) batch inhomogeneity and instability of the CRM in the uncertainty of the property values, and the determination of these uncertainty contributions.

Although this Guide has been developed to support best practice in the production and characterization of RMs, using it without carefully considering whether specific parts are applicable to the particular CRM may still cause its property values (and their uncertainties) to be established on a wrong or faulty basis. A user of this type of documentation should consider that it cannot substitute for “critical thinking, intellectual honesty and professional skill” (GUM:1993, 3.4.8). The quality of the “product” CRM depends as much on these aspects as on the use of proper procedures and methods.

Thorough knowledge of the material and its properties, and of the measurement methods used during homogeneity testing, stability testing and characterization of the material, along with a thorough knowledge of the statistical methods, are needed for correct processing and interpretation of experimental data in a typical certification project. It is the combination of these required skills that makes the production and certification of RMs so complex. The greatest challenge in these projects is to combine these skills to allow a smooth implementation of the project plan.

Most of the contents of this Guide can be applicable to the production of RMs. Requirements such as the traceability of the property values, the necessity of a full evaluation of measurement uncertainty, among others, apply to most categories of RMs to serve, for example, as calibrants or as a means to check the performance of a method, or to assign a value to another material.

Pharmacopoeial standards and substances are established and distributed by pharmacopoeial authorities following the general principles of this Guide. Specific guidance for the production of these kinds of RMs exists. It should be noted, however, that a different approach is used by the pharmacopoeial authorities to give the user the information provided by certificates of analysis and expiration dates. Also, the uncertainty of their assigned values is not stated since it is not permitted by the prescribed use of these RMs in the relevant compendia.

Reference materials — General and statistical principles for certification

1 Scope

This Guide gives statistical principles to assist in the understanding and development of valid methods to assign values to properties of a reference material, including the evaluation of their associated uncertainty, and establish their metrological traceability. Reference materials (RMs) that undergo all steps described in this Guide are usually accompanied by a certificate and called a certified reference material (CRM). This Guide will be useful in establishing the full potential of CRMs as aids to ensure the comparability, accuracy and compatibility of measurement results on a national or international scale.

In order to be comparable across borders and over time, measurements need be traceable to appropriate and stated references. CRMs play a key role in implementing the concept of traceability of measurement results in chemistry, biology and physics among other sciences dealing with materials and/or samples. Laboratories use these CRMs as readily accessible measurement standards to establish traceability of their measurement results to international standards. The property values carried by a CRM can be made traceable to SI units or other internationally agreed units during production. This Guide explains how methods can be developed that will lead to well established property values, which are made traceable to appropriate stated references. It covers a very wide range of materials (matrices), ranging from gas mixtures to biological materials, and a very wide range of properties, ranging from chemical composition to physical and immunoassay properties.

The approaches described in this Guide are not intended to be comprehensive in every respect of the production of an RM and the establishment of its property values, including the associated uncertainties. The approaches given in this Guide can be regarded as mainstream approaches for the production and value assignment of large groups of RMs, but appropriate amendments can be needed in a particular case. The statistical methods described exemplify the outlined approaches, and assume, e.g., normally distributed data. In particular when data are definitely not normally distributed, other statistical methods may be preferred to obtain valid property values and associated uncertainties. This Guide describes in general terms the design of projects to produce a CRM.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms*

ISO Guide 30, *Terms and definitions used in connection with reference materials*

Guide to the expression of uncertainty in measurement. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, 1993[1)]

1) This edition was corrected and reprinted in 1995.

International vocabulary of basic and general terms in metrology. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, 1993

NOTE The "Guide to the expression of uncertainty in measurement" will hereafter be referred to as "GUM", whereas the "International vocabulary of basic and general terms in metrology" will be referred to as "VIM".

3 Terms, definitions and symbols

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO Guide 30 and VIM, together with the following, apply. The symbols to be used are given in Clause 4.

3.1
reference material
RM
material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process

NOTE 1 RM is a generic term.

NOTE 2 Properties can be quantitative or qualitative (e.g. identity of substances or species).

NOTE 3 Uses can include the calibration of a measurement system, assessment of a measurement procedure, assigning values to other materials, and quality control.

NOTE 4 An RM can only be used for a single purpose in a given measurement.

3.2
certified reference material
CRM
reference material, characterized by a metrologically valid procedure for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability

NOTE 1 The concept of value includes qualitative attributes such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities.

NOTE 2 Metrologically valid procedures for the production and certification of reference materials are given in, among others, ISO Guide 34 and this Guide.

NOTE 3 ISO Guide 31 gives guidance on the contents of certificates.

3.3
property value
〈of a reference material〉 value attributed to a quantity representing a physical, chemical or biological property of a (certified) reference material

3.4
characterization
〈of a reference material〉 process of determining the property values of a reference material, as part of the certification process

NOTE 1 The characterization process provides the values for the properties to be quantified.

NOTE 2 In batch certifications, the characterization refers to the property values of the batch.

3.5
between-bottle homogeneity
bottle-to-bottle variation of a property of a reference material

NOTE It is understood that the term "between-bottle homogeneity" applies to other types of packages (e.g. vials) and other physical shapes and test pieces.

3.6
within-bottle homogeneity
variation within one bottle of a property of a reference material

3.7
blending
mixing of two or more matrix materials to obtain a material with specific properties

3.8
matrix material
material as sampled from nature, industrial production or elsewhere

EXAMPLES Soil, drinking water, air.

3.9
spiking
adding a known amount of a compound or element to a matrix material

3.10
short-term stability
stability of a property of a reference material during transport under specified transport conditions

3.11
long-term stability
stability of a property of a reference material under specified storage conditions at the CRM-producer

3.12
life time
⟨of a reference material⟩ time interval during which a reference material may be used

3.13
shelf life
⟨of an RM/CRM⟩ time interval during which the producer of the CRM warrants its stability

NOTE The shelf life is equivalent to the period of validity of the certificate, as described in ISO Guide 31.

4 Symbols

A_i	bias term (ANOVA)
a	number of groups (ANOVA)
B_i	bias term (ANOVA)
b	number of subgroups (ANOVA)
ε	error term (ANOVA)[2]
k	coverage factor
MS	mean square (ANOVA)
n	number of observations
n_0	(effective) number of (sub)group members (ANOVA)

[2] Throughout this Guide, the term error is used in the strict statistical sense, that is the difference between an observed value and its mathematical expectation.

p	number of laboratories in a collaborative study
s_{bb}	between-bottle (in)homogeneity standard deviation
s_{lor}	standard deviation due to lack of repeatability
s_{lts}	long-term (in)stability standard deviation
s_r	repeatability standard deviation
s_{stab}	standard deviation, due to (in)stability
s_{sts}	short-term (in)stability standard deviation
s_{wb}	within-bottle standard deviation
SS	sum of squares (ANOVA)
u_{bb}	standard uncertainty due to between-bottle (in)homogeneity
u_{char}	standard uncertainty due to characterization
u_{CRM}	standard uncertainty of a property value
u_{lts}	standard uncertainty due to long-term (in)stability
u_{sts}	standard uncertainty due to short-term (in)stability
U_{CRM}	expanded uncertainty of a property value
x_{char}	property value as obtained from characterization
x_{CRM}	property value of a CRM
δx_{bb}	error term denoting between-bottle (in)homogeneity
δx_{lts}	error term denoting long-term (in)stability
δx_{sts}	error term denoting short-term (in)stability
x_{ij}	result of a single measurement in the experiment (ANOVA)
μ	population mean (expectation)

NOTE 1 In some clauses, symbols are used to illustrate typical approaches to solve statistical issues in certification projects. These are explained in the text.

NOTE 2 The symbols MS and SS have been adopted from literature, and do not conform the ISO rules with respect to the use of symbols. For clarity however, it is felt that the convention in the scientific literature should prevail.

5 Design of a certification project

5.1 General

The production of a CRM requires a great deal of planning prior to undertaking any actual activity in the project. A substantial part of the planning deals with the amount of material needed, as well as with the design of the homogeneity, stability and characterization studies. The design also includes the choice of appropriate measurement methods for these studies. The number of samples to be produced is a very important variable in the planning process. The number of samples and the amount of raw material depend on all these factors. In the clauses about homogeneity testing (Clause 7), stability testing (Clause 8), and characterization (Clauses 9 and 10), guidance will be provided on how to plan and implement these processes as part of the certification project. A feasibility study may also be part of the project plan.

5.2 Project definition

The planning of a project starts with the definition of what CRM is to be produced. A typical example of such a definition reads as follows:

> “preparation of a soil CRM containing a series of trace elements at relevant content levels for environmental analytical chemistry with an uncertainty associated with the certified values of less than or equal to x %”

This definition covers the project quite well. What is relevant for environmental chemistry may differ from case to case, but it sufficiently narrows the range of materials. Likewise, “soil” also narrows the number of options for the matrix. In all cases it is important to specify what is to be produced. During the design stage of the project, the definition can be specified in more detail. Finally, the target uncertainty specified ensures that the material will be fit for its intended use. For example, uncertainties associated with values of calibration standards should be considerably smaller than uncertainties associated with values of materials for validation of trace environmental analytical methods.

The proper choice of the “stated references” whereto traceability of the property values is established is a major design issue; it strongly depends on what references are available, what is necessary in order for the particular CRM to serve the laboratories performing these measurements routinely, and what is technically feasible. As CRMs are primarily used to make later measurements traceable, the choice of proper references is crucial to the value of the CRM produced, both metrologically and commercially.

The scope for which the CRM is to be used should be stated as well. In most cases, the scope of use is implied by the project definition, but sometimes it needs further elaboration. Such a scope of use does not necessarily exclude other uses, but it should be kept in mind that such uses are not (necessarily) covered by the certificate or documentation provided. The scope for which the RM is to be used can be based on legislation and/or international treaties.

5.3 Transport issues

Prior to starting the actual work, it is important to consider whether the CRM, once ready, can be shipped in agreement with existing regulations. Many CRMs impose a risk with regard to health or safety when people are exposed to the material directly. Proper packaging and appropriate labelling are primary requirements to meet regulations for the transport of (potentially) hazardous goods. Sometimes, legislation or regulations prohibit the transport of materials having certain properties (e.g. viruses, diseases), which may mean that a CRM cannot be sold at all. It is recommended to review all aspects of transport and packing prior to starting the actual certification project.

5.4 Collection of starting material

The first task in a certification project is to obtain a sufficient amount of starting material(s) with the desired properties. For matrix materials, it should be noted that there may be restrictions with respect to the properties of materials. Some material/property combinations are rare, or may be rare in combination with other properties. Often a compromise must be found. In some cases, blending and/or spiking techniques may solve this problem.

The amount of material needed is dictated by the following:

- the number of samples of the (C)RM needed;
- the need for a feasibility study;
- the number of samples needed for the homogeneity study;
- the number of samples needed for the stability study;
- the number of samples needed for the characterization of the candidate CRM;
- the amount of material needed for one measurement.

The required number of samples needed of a candidate CRM is a commercial issue and should be carefully planned beforehand. An important variable is the number of samples likely to be distributed during the lifetime of the CRM. As lifetime is a function of intrinsic stability, this variable also affects the amount of raw material that is needed. For instance, many microbiological materials have limited intrinsic stability and, therefore, their lifetimes are expected to be shorter than, for example, that of a dry sediment certified for trace elements. For an equal number of samples to be dispatched per year, the number of samples needed for the microbiological material is smaller than for the dry sediment. On the other hand for microbiological CRMs, many more samples might be needed for stability testing in the first year(s), or through the entire lifetime of the material.

5.5 Feasibility study

When there are concerns about the feasibility of producing and characterizing a sufficiently homogeneous and stable CRM, a feasibility study may be considered (see Reference [11]). Questions with respect to, for example, the best way of preparing the sample, the stability of the material, or the fitness for purpose, may justify the inclusion of a feasibility study in the project (see References [11], [12]). Sometimes a feasibility study is organized to enable laboratories likely to be involved in the characterization to fine-tune their equipment and their procedures. For a feasibility study aiming at the characterization, it is recommended to have a batch of material slightly different from material used for the candidate CRM.

5.6 Required lifetime and shelf life

The expected lifetime of a reference material is an important variable in the planning of the certification project. Another relevant parameter with respect to the stability is the shelf life of the CRM. Depending on the nature of the mechanisms affecting the stability of the material, various actions may be taken to improve the shelf life and/or lifetime. Adjusting the water activity is one of the first options to be considered, as excessive drying or too high a water content can destabilize the material. In many cases, moisture plays a key role in mechanisms leading to instability of the matrix and/or parameters. In other cases, sterilization or pasteurization of the material might be considered in order to stop bacterial activity. However, these measures can also have a negative effect on stability. Relevant information regarding stability and storage conditions can be found in the literature or can be obtained from users of similar types of materials (industry, etc.). When preparing solutions, additives may increase the shelf life and/or lifetime. The shelf life of a material is a function of the storage conditions as well as a function of the quality of the stability study. The latter determines to what extent the results can be extrapolated (see 8.5).

5.7 Sample preparation

5.7.1 Preamble

It is difficult to give general guidance on the preparation of reference materials. This subclause is intended to give guidance on some specific aspects without the aim of being exhaustive. It is merely a collection of aspects needing careful consideration, which are frequently highly relevant for the success of a certification project.

5.7.2 Synthetic materials

Synthetic reference materials, such as pure substances, solutions and gas mixtures, are prepared in a completely different way from most matrix reference materials. For the preparation of pure substances, purification techniques may be necessary to reduce the total amount of impurities. The choice of these techniques depends on the main component of interest, and may include distillation and/or recrystallization techniques. After a subdivision process (when preparing a batch CRM), the batch should be treated as described in 5.7 to 5.9.

Many solutions and gas mixtures are prepared by means of gravimetry, for which a well-established uncertainty budget can often be obtained. The purity (or composition) of the starting materials enters into the model for calculating the composition of the candidate CRM, as does its uncertainty. For the preparation of batches of materials, volumetric techniques are widely used as well. Usually volumetric methods are somewhat easier to handle, but they are usually also associated with a larger uncertainty than would be the case when prepared gravimetrically.

5.7.3 Blending of materials

Blending of two or more matrices may be considered if a particular property value is considered to be too high or too low. The process is best carried out with matrices of similar kinds, although what is considered to be "similar kinds" may differ widely. For proper blending, the material should be in such a state that agglomeration of particles is suppressed. Usually the moisture content of the materials involved is the dominant factor. If the material is "air-dry", usually (but certainly not always) agglomerates disappear during a good mixing process. The same is true for materials that behave like slurries. There is a potential problem when the agglomerates do not disappear during mixing. Some level of agglomeration of particles may be inevitable. For instance, soy powder with less than 2 % water is still sticky.

A further requirement for proper blending of different matrices is that the densities and the particle size distributions of the materials being blended be sufficiently similar and, for the distribution, sufficiently narrow. This will substantially reduce the segregation risks. With appropriate technology and correct implementation of particle size reduction and blending techniques, it is usually possible to obtain a batch of material that has good properties with respect to homogeneity and stability.

In case of doubt, the blended material may be subjected to a quick homogeneity test, where several portions of the blended material are investigated for homogeneity of the properties to be certified. Such a study may be run on a small number of portions, but a large enough number to obtain some idea about the homogeneity. Typically, 10 portions should be considered to provide meaningful results for taking a decision as to whether the blend material is suited for further processing.

5.7.4 Spiking

There are cases where spiking should be considered as a suitable method for the production of a reference material. Such cases include extracts prepared from solid-state materials. Another example is a series of three CRMs of PCBs in pork fat, where the CRM at elevated temperatures is a liquid. Other examples where spiking is a good method for obtaining CRMs of desired properties are liquids, metals and alloys, oils and workplace atmospheres.

A major problem with spiking is the achievement of sufficient homogeneity and stability of the candidate reference material. Using a proper spiking method can lead to a material that fulfils the requirements with respect to homogeneity and stability, even for solid-state materials. A suitable spiking method for solids is, for example, an "incipient wetness" technique, where the component to be spiked is dissolved in a suitable amount of solvent that is just sufficient to completely wet the surface of the solid. The solvent should be chosen in such a way that its rate of evaporation can be controlled. If the rate of evaporation is too high, the spike may come out again from the pores and cluster. In that case the spike will not be sufficiently well bonded to the surface, which has an impact on the stability of the material. Too low rates of evaporation will lead to migration of other constituents present in the matrix, or even their loss.

For some groups of matrix CRMs, however, spiking is clearly an inappropriate method for obtaining a material with desired values for properties to be certified, as it may lead to CRMs that behave completely differently from normal routine samples. As a rule, major differences in the binding of the naturally incurred and spiked analytes can be expected, leading to differences in, for example, the extraction behaviour. The equivalence of the spiked material to naturally (contaminated) material should therefore be checked to make the material representative of real samples.

5.7.5 Homogenization and subdividing

The sampled material usually undergoes several preparation steps before it becomes a reference material. Necessary steps in this process include drying, particle size reduction, sieving, stabilization and subdividing/bottling. At the design stage of the project, it should be established how far the sample preparation will be extended. For instance, it is possible to prepare a sampled material in such a way that it can be measured directly as an extract. In many cases, however, it is preferable that the sample preparation should leave the sampled material in its original state, although heterogeneity should usually be decreased and stability should be increased as a result of the sample preparation process.

The required uncertainty of the property values of the RM and the required lifetime set requirements with respect to the choice of sample preparation techniques. It should be borne in mind that the way in which the candidate reference material is prepared influences the possible use of the material. For example, distributing an extract will make it impossible to check for the accuracy of the extraction step in the customer's laboratory. Therefore, the objectives of preparing a CRM should be kept in mind when deciding how the raw material is prepared to become suitable to be certified in view of the scope of use of the CRM.

5.8 Homogeneity study

A homogeneity study is necessary in batch certification projects to demonstrate that the batch of bottles (units) is sufficiently homogeneous. Aspects of quality assurance are as important as the determination of the remaining batch between-bottle variation, which is an uncertainty component to be included in the uncertainty estimate of the property value of the CRM. Even when a material is expected to be homogeneous, as in the case of solutions, an assessment of the between-bottle inhomogeneity is required. When dealing with solid-state reference materials, including slurries and sludges, a within-bottle homogeneity study should be foreseen to determine the minimum sample intake. In principle, this homogeneity study does not add to the uncertainty of the property value in question. The number of extra samples needed mainly depends on the between-bottle homogeneity study. The minimum number of bottles selected at random is between 10 and 30, but should generally not be smaller than 10.

The optimal number of samples for a homogeneity study can be determined by statistically supported design techniques. Such methods usually take into consideration the inability of detecting any inhomogeneity, for example due to the uncertainty of the measurements. Furthermore, the number of bottles depends on the batch size, so that the number of samples picked from the batch may be considered to be "representative" of the whole batch. This requirement should be balanced with the uncertainty of the measurements, which is (under repeatability conditions) a function of the repeatability standard deviation of the measurement and the number of replicates. The above-mentioned statistical techniques may be of help with balancing the number of bottles and the number of replicates, so that the best approach is chosen.

5.9 Stability study

Stability testing aims to determine the remaining degree of instability of the candidate RM after preparation, or to confirm the stability of the material. Even "stable" materials may show instability for one or more property values. A distinction is made between the stability under specified

- storage conditions (long-term stability), and
- transport conditions.

As in the case of a homogeneity study, quality assurance aspects are as important as determining the uncertainty budget due to instability effects. The long-term stability concerns the remaining instability of property values of the CRM under specified storage conditions. It is therefore important to specify these conditions accordingly and to study the stability of the material under the same conditions. A reference temperature should be chosen such that it is practically certain that the material is stable at that temperature. Many biological and environmental reference materials show some degree of instability, despite the effort put into defining/determining optimal storage conditions. Transport conditions should ideally be chosen so that the instability of the material during transport does not exceed that of the material on the shelves of the producer. Short-term stability is therefore only relevant as an uncertainty component when the stability of a CRM is affected by the specified transport conditions (e.g. from the producer to the user) in excess of the storage conditions.

The short-term stability study is typically carried out at different temperatures, to study the effect of different temperatures on the properties of the material. Temperatures of samples can vary during transport between –50 °C up to +70 °C, depending on the type of packaging and transport modality. Based on the observed effects, the transport conditions may be defined and packaging instructions drawn up to effectively eliminate any unwanted side effects. A short-term stability study takes typically 1 to 2 months, but may be extended when the optimal storage conditions are to be determined simultaneously.

The stability study requires a considerable number of bottles (units). For each point in time, more than one bottle should preferably be available. As most long-term stability studies last between 24 and 36 months, with typically 5 or 6 points in time, at least 10 to 12 bottles are needed per temperature. When the design foresees multiple temperatures, the number of bottles should be multiplied accordingly. For a short-term stability study, usually 3 to 5 points in time are used, over 2 weeks. Following the same reasoning as for a long-term stability study, the number of bottles should be between 6 and 10 for a short-term stability study per temperature. The inhomogeneity of the material will also influence the number of units needed for the test of stability. If the material is inhomogeneous, it is advantageous to make single determinations on several bottles rather than replicate determinations on a few bottles.

The preferred method for conducting a stability study in a batch certification is to work under repeatability conditions. Otherwise, the estimated uncertainty due to instability is unnecessarily enlarged due to the reproducibility effects in the results during stability testing. Working under repeatability conditions is possible using the isochronous design (see Reference [13]). All samples are kept at a reference temperature at which it is assumed that no instability is encountered (not necessarily the envisaged storage temperature). The samples are subjected to the temperature under test in the stability study and kept at this temperature until all samples have been measured. The points in time are defined by the time elapsed between the moment that they are put at the temperature under test and the moment that they are measured.

For the classical layout (see 8.2), a measurement method should be chosen with good reproducibility properties. Because maintaining good reproducibility of a measurement method is considerably harder than maintaining good repeatability during a single experiment, the isochronous design is advantageous over the classical design. Apart from this aspect, the uncertainty in the assessment using the classical design is in any case greater than in the isochronous case, which means that the shelf life that can be derived from an isochronous stability study will (for a given level of uncertainty) be longer than for a stability study using the classical layout. These advantages compensate well for the disadvantage of having no data during the stability study, in particular for methods with (relatively) poor repeatability and reproducibility. When intermediate data are required, those measurements should be taken independently from the isochronous stability study. When certifying a single artefact, there is no choice but the classical layout.

The experimental design for a stability study, including determination of the optimal number of points in time and the number of samples included, may be based on a statistical design, appreciating, for example, the inability of the measurement method to detect any instability. Furthermore, an empirical model is used as in most stability studies, so the number of points in time should be sufficiently large that a proper assessment of the validity of the model can be carried out. For a linear model for example, which has two parameters (intercept and slope), at least 3 or 4 points are needed, but often more to make a more accurate assessment. For models with more parameters, the number of points (in time) in the stability study should be increased accordingly.

5.10 Choice of measurement methods

The measurement method used for the homogeneity study should have very good repeatability and selectivity. For a stability study where samples are measured on different days, the selectivity and the reproducibility of the measurement method are of primary importance. Therefore, methods for homogeneity and stability studies are not necessarily the same. This is not a problem so long as traceability of the results of the homogeneity and stability studies and characterization to a common reference are established. Such a reference may be a material that is suitable for assessing the various calibrations or results from different measurement methods. Ensuring the traceability of all measurements in a certification project is an important requirement (see, for example, ISO Guide 34 and Reference [14]).

For the characterization of the candidate reference material, especially in the case of matrix reference materials, it is often highly desirable to use multiple methods, and often also multiple laboratories. Both the methods and the performance of the laboratories should represent “state-of-the-art”, and they should be able to make their measurements traceable to the references specified in the design of the project.

The characterization of a candidate reference material may take place in different ways. There are two mainstream approaches:

a) characterization by a single method, and

b) characterization by multiple methods and/or multiple laboratories.

Approach b) includes experimental set-ups known as a collaborative study or a collaborative trial. Both names underpin the joint effort of the coordinator and participants to characterize the reference material. In all cases, the measurement procedures used in the characterization should be made traceable to “stated references”, preferably to SI units. The aspect of traceability of measurement results goes well beyond the actual measurements; it also includes the transformation of the sample. Transformation of a sample means bringing the material (or artefact) from one (physical, chemical) state to another. Examples of such transformations of a sample include the destruction of the sample, and the extraction of the species to be measured.

Finally, the measurements of the homogeneity study, stability study and the characterization of the material should be combined in order to obtain a proper estimate for the property value and its standard uncertainty. A requirement for the data is that they refer to the same “scale”; that is, that all measurements are carried out with properly calibrated equipment, and that the results of these calibrations can be compared one to another. In particular, when more laboratories are involved, the use of some kind of standard substance, mixture or solution may be used to verify the degree of agreement between calibrations. This aspect is partly an issue of defining proper references and thus establishing metrological traceability, partly of being able to demonstrate the validity and comparability of the results obtained in the various stages of the project.

5.11 Certification

The certification of a CRM is described in Clause 11.

5.12 Summary of project design

In summary, the preparation of reference material involves the following steps:

a) definition of the reference material, i.e. the matrix, the properties to be certified and their desired levels, and the level of uncertainty desired;

b) design of a sampling procedure;

c) design of a sample preparation procedure;

d) selection of measurement methods appropriate for homogeneity and stability testing;

e) design of the characterization of the reference material;

f) sampling;

g) sample preparation;

h) choice of suitable methods for the characterization;

i) homogeneity testing;

j) stability testing;

k) characterization of the reference material;

l) combination of the results from homogeneity testing, stability testing and characterization, including a full evaluation of measurement uncertainty;

m) design of a certificate and, if appropriate, a certification report.

6 Evaluating measurement uncertainty

6.1 Basis for evaluating the uncertainty of a property value of a (C)RM

The basis for any evaluation of measurement uncertainty is the GUM. Most of the projects that lead to a CRM can be evaluated through the approach as given in Clause 8 of GUM:1993. For a CRM, this procedure can be summarized as follows:

a) Express the relationship between the property value to be certified and all input quantities on which the property value depends mathematically. The relationship should include all quantities that might contribute significantly to the uncertainty of the property value, and is called the measurement model.

b) Determine the values for all input quantities, based either on statistical analysis of a series of data or by other means.

c) Evaluate the standard uncertainty for all input quantities using a type A evaluation for quantities obtained from a statistical analysis of data, or using a type B evaluation for all other quantities.

d) Evaluate the covariances between any input quantities.

e) Calculate the property value (x), i.e. the value of the characteristic to be certified.

f) Determine the combined standard uncertainty associated with the property value from the standard uncertainties and covariances associated with the input quantities, using the propagation formula as given in Clause 5 of GUM:1993.

g) Determine a coverage factor k to obtain an expanded uncertainty U, for which it may be assumed that the interval $[x - U, x + U]$ contains a large fraction of the distribution of values that could reasonably be attributed to the characteristic being certified. The choice of a coverage factor should be based on the required level of confidence (often 95 %), the probability density function of y and (if applicable) the number of degrees of freedom.

h) The property value should be reported together with the expanded uncertainty U and the coverage factor k, following the recommendations of ISO Guide 31.

In the vast majority of the cases, the approach as described can be followed. There are however some situations, where other approaches should be chosen, as discussed in the GUM. Such situations include

— cases where there is no closed mathematical form for the model describing the relationship between the property value and the input quantities, and

— cases where the linear approximation, as obtained by applying the formula for combining and propagating uncertainties, is clearly invalid.

Other statistical techniques, including Monte Carlo or bootstrap methods, may be used to determine the uncertainty associated with the property value of a CRM in these cases. For the purposes of this Guide, it is assumed that the approach as outlined can be followed. Other cases should be dealt with in agreement with the GUM.

The details on how to evaluate single components of uncertainty are covered by the GUM. In many cases, aggregated uncertainty components can be defined in order to take advantage of existing data, such as the results of validation studies, as described in Chapter 7 of Reference [15].

Uncertainty components that need specific guidance on their evaluation include the uncertainty due to batch inhomogeneity, long-term stability and short-term stability issues. The evaluation of these uncertainty components is covered in this Guide in Clauses 7 (homogeneity testing) and 8 (stability testing). Some additional guidance is given in Clauses 9 and 10 on the evaluation of measurement uncertainty of the determination of the property value for the batch.

The uncertainty of property values from single-artefact CRMs that are certified based on a single calibration may be carried out using the normal procedures as outlined in the GUM. It should be noted, however, that the uncertainty budget of this type of CRM should also include long-term stability effects.

6.2 Basic model for a batch characterization

Modelling a characterization process for the evaluation of uncertainty is neither a routine task nor a strictly mathematical one. The establishment of a proper model for a property value of a specific candidate CRM is a complex task, which should be carried out with great care to account for all relevant details of the procedures followed to produce and certify the material. One of the basic requirements of the model is that all factors are included that could significantly contribute to the uncertainty associated with the property values of the CRM. Therefore, in order to be complete, the combined standard uncertainty on a reference material should acknowledge that homogeneity and both long- and short-term stability also play an important role in addition to the characterization of the batch. Therefore, the uncertainty of a reference material can be expressed as follows:

a) uncertainty of the certified value as obtained for the batch (characterization);

b) transferred to a single package (homogeneity);

c) as dispatched to the customer (short-term stability);

d) at the time of sale (long-term stability).

This definition of the uncertainty associated with a property value of a CRM considers the following factors:

e) the (thorough) characterization of the material is accompanied by an uncertainty;

f) the user will (as a rule) use only one sample at a time;

g) the material will be stored for a longer period of time by the producer/seller;

h) the material must be transported to the user.

These factors could all significantly contribute to the uncertainty associated with the value assigned to the measurand (i.e. the value to be certified) for a CRM. Appreciation of these influencing factors does not go any further than regular operations. Uncertainty evaluation is not intended, and should not be used, to account for accidents, mistakes, improper use, improper transport, etc., of the CRM. This approach is in agreement with GUM:1993, 3.4.8.

The model can be expressed as follows:

$$x_{\mathrm{CRM}} = x_{\mathrm{char}} + \delta x_{\mathrm{bb}} + \delta x_{\mathrm{lts}} + \delta x_{\mathrm{sts}} \tag{1}$$

where

x_{CRM} denotes the property value;

x_{char} denotes the property value obtained from the characterization of the batch or, in the case of a single artefact characterization, the property value obtained for this artefact;

δx_{bb} denotes an error term due to the between-bottle variation;

δx_{lts} and δx_{sts} are error terms due to the long-term and short-term instability.

Usually, the homogeneity and stability studies are designed in such a way that the values of these error terms are zero, but their uncertainties are not.

Assuming independence of the variables, the uncertainty associated with a property value of a CRM can be expressed as

$$u_{\mathrm{CRM}} = \sqrt{u_{\mathrm{char}}^2 + u_{\mathrm{bb}}^2 + u_{\mathrm{lts}}^2 + u_{\mathrm{sts}}^2} \qquad (2)$$

using the error propagation formula in GUM:1993, E.8, the uncertainty components u_{bb} (between-bottle standard uncertainty), u_{lts} (long-term stability standard uncertainty) and u_{sts} (short-term stability standard uncertainty) correspond to the error terms in the model. The combined standard uncertainty associated with the property value of the CRM may be related to the shelf life of the material (see Clause 8).

Sometimes, the long-term stability term is a function of time, such as for reference materials certified for radioactive isotopes. The model used for evaluating the uncertainty associated with the property value of this type of CRM should account accordingly for the time dependence of the certified value.

General recommendations on modelling measurements can be found in the GUM and in several supplementary documents, for example Reference [15]. Some specific guidance with respect to modelling and data evaluation is given in Clauses 7 to10 with respect to modelling key steps in the certification of reference materials.

Under some circumstances, it is possible to deviate from the basic model as stated. Such circumstances include situations where no transport of samples takes place, or where it is explicitly stated that the uncertainty on the certificate does not include the transport of the samples. Specific guidance on the estimation of these uncertainty components is given in Clauses 7 (homogeneity testing), 8 (stability testing), 9 and 10 (determination of the property value).

If, for example, a sample that is sensitive to elevated temperatures travels for 6 weeks from the producer to a customer, where the producer foresees only 1 week at maximum, the properties of the CRM may have undergone severe changes. Provided that the producer specifies on his certificate or, if deemed more appropriate, in additional documentation, the producer is entitled to limit the short time stability study accordingly.

6.3 Uncertainty sources

Apart from the uncertainty sources mentioned, the uncertainty sources commonly encountered in measurement procedures, should also be included in the model. Both the GUM and Reference [15] list these uncertainty sources. When constructing a model, it is recommended to follow such a generic list in order to reduce the effort of getting all relevant uncertainty components. Often, the measurement methods have already been evaluated in terms of measurement uncertainty, and the models available for these methods can be applied for the evaluation of the uncertainty of a property value of a CRM as well. It should be noted that any change in a particular measurement procedure should be accompanied by a review of the uncertainty model.

Often, uncertainty models of measurement methods contain aggregated components, i.e. uncertainty components which depend on several others. These aggregated components may lead to covariances (see 6.1), even if these do not appear when the measurement method is used for a routine measurement. The evaluation of covariances and correlations is crucial to obtain a correct estimate of the combined standard uncertainty associated with the property value of a CRM. To facilitate the process of detecting covariances, it is recommended to document which uncertainty components are contained in the aggregated uncertainty components. This documentation allows relatively quick identification of possible sources of covariances and correlations. In GUM:1993, appendix F, some further guidance is given on how to evaluate the resulting covariances.

6.4 Issues with distribution functions

Most statistical techniques require implicitly or explicitly assumptions concerning the probability density function of the variable under study. The approach of the GUM does not form an exception, because somewhere in the evaluation process, the probability density function will have been determined or assumed. The models in use for the certification of reference materials do not form an exception to this principle, as they build forth on basic statistical theory. Frequently, these assumptions are made implicitly (e.g. by using a particular statistical estimation/type B evaluation technique) rather than explicitly. Many statistical methods assume, for example, normally distributed data. This assumption is also underlying most of the statistics in this Guide. For most data from composition measurements, this approximation is fair, whereas for other measurements, such as counting of small numbers, this assumption may be invalid.

Regression analysis and analysis of variance are based on the assumption of normally distributed data. Nevertheless, these statistical tools work well on data that have a unimodal distribution function, as long as it is used to estimate variances (see, for example, Clauses 7 and 8, and Reference [21]).

A notorious problem arises when the experimental distribution of characterization data shows multiple maximums. In the worst case, this means that the material cannot be certified due to lack of agreement among results from laboratories and/or measurement methods. Assigning a single property value is only useful when there is agreement among methods and/or laboratories. Small discrepancies may be resolved by introducing an additional uncertainty component, appreciating this effect. If there is agreement between the laboratories using a particular method, then a method-dependent certification may be considered, resulting in method-dependent property values. If there is no agreement between the laboratories and grouping by methods does not solve the problem, then the characterization data are unsuitable for establishing property values.

6.5 Use of ratios

A potential problem exists in using ratios, for instance in stability studies [16]. It should be noted that the ratio of two normally distributed variables is not necessarily also normally distributed [17]. The actual distribution of the ratio of two variables depends on both distributions of the two variables involved, as well as on the actual values of the parameters of these distribution functions. In particular when a property value can be zero, as for instance with some colour measurements, the use of ratios may lead to problems, as the ratios will follow the Cauchy distribution [17]. This distribution has no moments, which means that, for example, the variance is not defined. As a consequence, it is not possible to evaluate the measurement uncertainty on the basis of the assigned probability distribution.

The minimization of the effect of random error components due to measurement can also be obtained through correct application of the law of propagation of uncertainty; i.e. the inclusion of the necessary covariance terms between the two variables forming the ratio. This “reduction of uncertainty” is often the desired effect to be achieved by using ratios [16]; using the law of propagation of uncertainty on the observed data has the advantage of being safe with respect to artefacts in the distribution function of the ratios and, at the same time, leads to the desired effect of “suppressing random error components” [18].

6.6 Choice of a coverage factor

The coverage factor used in step g) of the approach outlined in 6.1 is determined on the basis of the distribution function assumed for the property value (often the normal distribution) and the level of confidence (often 95 %). As a consequence, a coverage factor $k = 2$ is often assigned on this basis (normal distribution, 95 % level of confidence). When the (effective) number of degrees of freedom is considered to be low, Student’s t-distribution may be used instead to assign a coverage factor.

In cases where the assigned distribution of the property value is considered to be asymmetrical, such as in the case of the result of a count following the Poisson distribution, a confidence interval should be stated rather than the expanded uncertainty and a coverage factor.

6.7 Recertification

Over time, the actual property value of a CRM may drift from the certified one. When the actual property value of a CRM lies outside the range indicated on its certificate, there are two basic ways to deal with the problem:

— to withdraw the CRM/RM, or

— to do a recertification of the material.

The choice between the two options is based on both economical and technical factors. Technical factors for withdrawal may include, for example, a deterioration of the matrix or one or more of its constituents, which might be the conclusion drawn from a stability test or a stability monitoring (see 8.4). A withdrawal of a CRM from the market may be preferred when the remaining batch of items has become too small to do a recertification.

A recertification implies that (relevant parts of) the homogeneity test, the stability test(s), and/or characterization of a reference material are carried out again. Improvement of the measurement capability in a particular field may also be a reason for a recertification (when economically feasible). The material as produced may still be good enough, however the establishment of the property values should be improved to reduce their uncertainty to become useful again for the users of the CRM.

Another type of recertification observed in practice is due to a gradual change of a property of a material. An example of this type is the calorific value of coal, which changes over time even when coal is stored under the best possible conditions.

7 Homogeneity study

7.1 Preamble

Most RMs are prepared as batches of items (e.g. bottles, vials or test pieces). The final step in the preparation of many RMs is the subdivision into usable items. A subset of the batch of items, typically 10 to 30, is chosen by a sampling scheme to undergo a homogeneity study. There are various methods for selecting the subset from the batch (e.g. random sampling, stratified random sampling or systematic sampling). Random sampling or stratified random sampling schemes are mostly used in practice and usually provide a subset that can be regarded as representative of the whole batch. If it is certain that inhomogeneity will not be detected in the batch, systematic sampling schemes may be used too.

The results of the between-bottle homogeneity study [3] provide for the evaluation of one of the uncertainty components in the certification model (see Clause 6). The magnitude of this uncertainty component can vary widely, mainly depending on the nature of the RM. This type of homogeneity testing is only applicable when a certificate valid for a batch of items is issued.

A second important type of inhomogeneity is within-bottle homogeneity, the impact of which may be reduced considerably by providing proper instructions for use. These instructions may include remixing of the sample and, for granular materials, a minimum sample intake. This is the smallest test portion which, when drawn correctly, may be considered as being representative of the RM within the certified uncertainty.

7.2 Materials

RMs prepared as solutions or pure compounds (if certified for purity; not for impurities) are expected to have a high degree of homogeneity on physical (thermodynamic) grounds. These materials can, however, also show some heterogeneity, for example due to a density gradient or metals containing occluded gases. The objective of the test for homogeneity for these materials is mainly to detect any impurities, interferences or irregularities that may be due to undetected problems during the preparation. In these cases, a very small if not negligible uncertainty contribution is expected from the between-bottle homogeneity study. Even in these cases, where perfect homogeneity may be assumed, such an assumption should be verified experimentally by a homogeneity study.

Materials such as mixed powders, ores, alloys, etc. are heterogeneous in composition by nature. RMs prepared from such materials should therefore be tested to assess the degree of inhomogeneity. The magnitude of the uncertainty component due to between-bottle inhomogeneity can still be small or even

[3] Where reference is made to “between-bottle homogeneity”, it is understood that the same applies to other physical shapes of an RM, such as vials or test pieces.

negligible in comparison with the uncertainty associated with, for example, stability testing or characterization, but in some cases it is inevitable that it is of the same magnitude as the uncertainty component from the determination of the property value (characterization). Much depends on the options available during preparation to reduce the batch inhomogeneity.

7.3 Concept of homogeneity

In theory, a material is perfectly homogeneous with respect to a given characteristic if there is no difference between the values of this characteristic from one part (item) to another. However, in practice, a material is accepted to be homogeneous with respect to a given characteristic if a difference between the values of this characteristic from one part (or item) to another is negligible when compared to the uncertainty component from, for example, characterization.

There is an experimental limit to the detection of batch inhomogeneity (u_{bb}). Care should be taken not to underestimate this uncertainty component due to limitations arising from, for example, the measurement method. In particular, when only methods with poor repeatability are available, such an underestimation risk exists. Furthermore, whenever possible, the subsamples taken for measurement should be sufficiently large, so that this type of subsampling does not contribute significantly to the uncertainty due to repeatability of measurement (see also 7.10).

This subclause deals mainly with bulk inhomogeneity, since for most reference materials this type of inhomogeneity is the most relevant one. There are important exceptions, however. For example, in surface analysis, reference materials may be wafers or foils. The relevant inhomogeneity is, of course, that across the surface, and not in the direction perpendicular to the surface. Most considerations in this clause can be valid for other types of inhomogeneity as well, but the guidance offered should be compared with more specialized literature, including International Standards describing the relevant measurement methods.

7.4 Practice

Ideally, an RM should be characterized with respect to the degree of inhomogeneity for each characteristic of interest. For RMs with a relatively large number of properties to be certified, the assessment of the degree of inhomogeneity for all characteristics may be burdensome both economically and physically and, in some cases, unfeasible. It should, however, be realized that the quality of the RM produced depends (among other things) on the correct assessment of the batch inhomogeneity.

In practice, the degree of homogeneity of such RMs may (under certain conditions) be assessed only for selected characteristics when the preferred approach is not feasible. It is essential that these characteristics be appropriately selected on the basis of established chemical or physical relationships; for example, an inter-element concomitance in the mineral phases of an RM makes reasonable the assumption that the RM also has a similar degree of homogeneity for the non-selected elements. Other examples where a reduction of the characteristics included in the homogeneity study is possible include hard and brown coal, where inhomogeneity is properly reflected by the ash and/or sulfur content. For other characteristics, the batch homogeneity is usually better than for these characteristics.

In all cases, additional evidence should be gained about the homogeneity of characteristics not covered experimentally by the homogeneity study. Such evidence can be gained, for instance, from literature, through the stability study, or the characterization of the material. The evidence thus gained should allow quantitative transfer of the magnitude of inhomogeneity observed for one characteristic to another, with sufficient evidence that the degree of inhomogeneity is not underestimated.

7.5 Measurements

Measurements in a homogeneity study should be carried out under repeatability conditions (see ISO 5725-1 [1] for a definition of repeatability conditions). Furthermore, the repeatability standard deviation of the measurement method should be small. If possible, a situation should be achieved where the uncertainty associated with the determination of a single bottle ($s_r/\sqrt{n}$) is considerably smaller than the (expected) combined standard uncertainty from the determination of the property value. In some cases this is not feasible, which might require an alternative approach to the one given in 7.9 [19], leading generally to a higher uncertainty estimate.

The measurements should be carried out in such a way that a trend (drift) in the measurements can be separated from a trend in the batch of samples. This can be achieved by measuring the replicates of the samples used in the homogeneity study in a randomized order. Alternatively, the order of measuring samples may also be reversed between the replicates, as in the following example.

EXAMPLE Suppose, 10 samples are used for a homogeneity study, with 3 replicates. A suitable scheme for conducting the measurements reads as follows:

Replicate No. 1:	1 – 3 – 5 –7 – 9 – 2 – 4 – 6 – 8 – 10
Replicate No. 2:	10 – 9 – 8 – 7 – 6 – 5 – 4 – 3 – 2 – 1
Replicate No. 3:	2 – 4 – 6 – 8 – 10 – 1 – 3 – 5 – 7 – 9

A trend due to drift in the measurements can be detected by performing a trend analysis on the results in the exact order of measurement. A trend due to the sample preparation can be detected by analysing the bottle averages as a function of their sequence number. It is therefore important that the sequence number of a sample batch is logically related to the sample preparation process and, in particular, the subdivision process.

7.6 Statistically valid sampling schemes and trend analysis

The sampling scheme, used to pick the bottles (items) for the homogeneity study may be random, random stratified or, in some cases, systematic. The sampling scheme should take into consideration potential weaknesses in the method of preparing samples, thus allowing a critical examination of the prepared batch. Stratification is recommended in many situations, since this guarantees that the bottles picked for the homogeneity study are equally distributed over the whole batch. Systematic schemes may be applied when there is practically no risk of overlooking systematic effects or trends in the batch.

The measurements should be carried out in such a way that any trends that might be present in the samples do not interfere with any trends that might be in the measurements themselves. In a measurement scheme, this can be achieved by, for example, randomizing the order of bottles in combination with changing the order in which the samples are measured.

Prior to determining the magnitude of the between-bottle homogeneity standard uncertainty, the experimental data obtained should be inspected for trends. In 8.3.1, a basic recipe for trend analysis is given in the context of stability studies, a methodology which can also be applied for the data of a homogeneity study as a function of the bottle number. If a significant trend in the bottles is present, then usually the batch produced is unsuited for a batch certification. A trend in the measurement results is something to correct for, irrespective of whether it is statistically significant or not. A method for trend analysis and, if necessary, for developing a correction for instrument drift, is the inclusion of a quality control sample that can be fed directly to the instrument. When observing a trend in the batch, a redesign of the subdivision procedure may be necessary to eliminate such a trend effectively.

7.7 Evaluating a homogeneity study

A basic model for a homogeneity study comprising $i = 1 \ldots a$ bottles and $j = 1 \ldots n_i$ measurements can be expressed as follows (see, e.g. References [20] to [22]):

$$x_{ij} = \mu + A_i + \varepsilon_{ij} \qquad (3)$$

where x_{ij} is the result of a single measurement in the homogeneity study; μ is the (mathematical) expectation of x_{ij}, which is the value that the grand mean (mean of means) takes up when the number of repeated measurements tends to infinity. If the measurements are unbiased, then μ is equal to the true value. The terms A_i and ε_{ij} are the error terms for the between-bottle homogeneity and the random measurement error. The variances of these terms are the between-bottle variance and the repeatability variance respectively. Usually, it can be assumed that A_i and ε_{ij} are mutually independent, that is, the between-bottle inhomogeneity does not influence the repeatability of measurement or vice versa. Furthermore, it can often be assumed that the variable A_i is normally distributed, with mean zero and variance $\sigma_A{}^2$. Likewise, it can often be assumed that the random measurement error ε_{ij} is normally distributed variables with mean zero and variance σ^2 [21].

Various experimental designs can be developed for a between-bottle homogeneity study. In B.2, a case using a fully nested one-way analysis of variance approach is described.

7.8 Between-bottle homogeneity study

A between-bottle homogeneity study aims to determine the between-bottle variation. The “groups”, as described in the previous clause, represent bottles (units). Two typical experimental set-ups for a between-bottle homogeneity study are visualized in Figures 1 and 2.

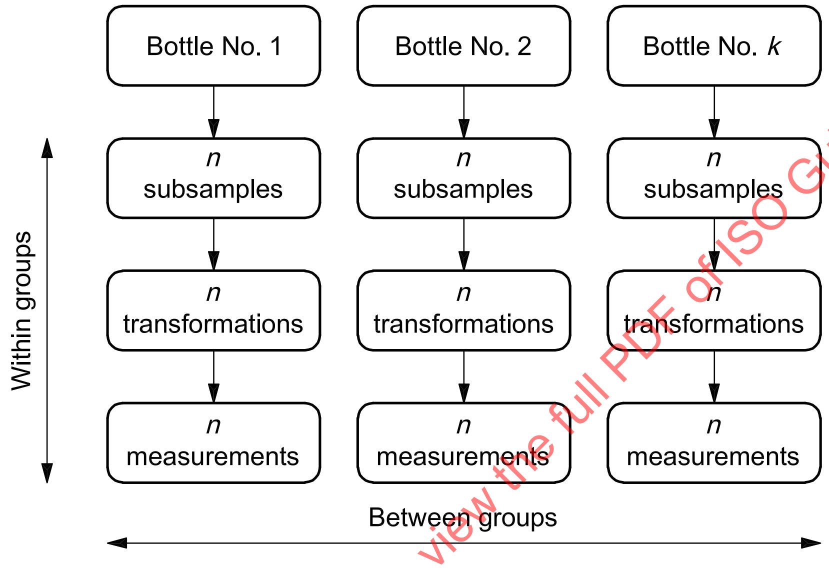


Figure 1 — Layout of a between-bottle homogeneity study
(from Reference [22])

In Figure 1, an ideal case is shown in which subsampling of the items is possible and has been carried out. In this design, because multiple test portions have been taken from each sample of the batch and individually transformed, the variance “between bottles” only includes the between-bottle heterogeneity, while the variance “within bottles” includes the uncertainty due to measurement, transformation and subsampling. From a perspective of obtaining an unbiased estimate of the heterogeneity of the material, this is the ideal situation.

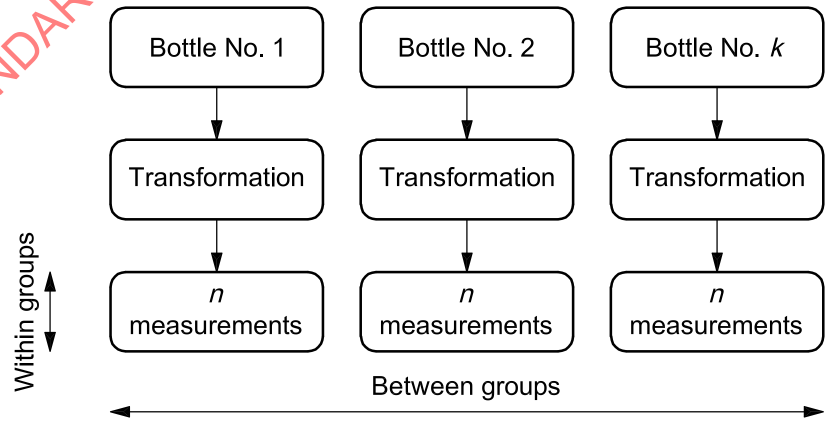


Figure 2 — Alternative layout of a between-bottle homogeneity study
(from Reference [22])

In Figure 2, a design is shown for the case where subsampling of the items is impossible or just not carried out, for example for economic reasons. In this design the effect of between-bottle homogeneity is included in the variance “between groups”, as are any effects arising from the transformation of the sample. The variance “within groups” covers only the repeatability of the measurement. With test pieces, or “single-shot” samples, often only one test is possible, so in this case, n, the number of replicates, equals 1. In these cases, there are no within-bottle homogeneity effects to account for. In those cases where the sample allows multiple measurements after transformation, n will generally be greater. In those cases where $n > 1$, the data can be treated with an analysis of variance (see A.1 and B.2).

If a one-way analysis of variance approach is used, then s_{bb} can be computed in both cases by

$$s_{\text{bb}}^2 = s_A^2 = \frac{MS_{\text{among}} - MS_{\text{within}}}{n_0} \qquad (4)$$

In these cases, the between-bottle variance $s_{\text{bb}}{}^2$ is identical to $u_{\text{bb}}{}^2$.

7.9 Insufficient repeatability of the measurement method

It is not always feasible to perform a homogeneity study with a measurement method that is sufficiently repeatable. In those cases, an alternative approach may be necessary that attempts to estimate the maximum effect. If u_{bb} denotes the uncertainty component due to batch inhomogeneity to be included in the model for certification, it should be noted that

$$\frac{MS_{\text{among}} - MS_{\text{within}}}{n} \le u_{\text{bb}}^2 \le s_{\text{bb}}^2 + \frac{s_r^2}{n_0} \qquad (5)$$

The repeatability variance can be derived separately, or be set equal to MS_{within}. The right-hand side of this expression represents the squared standard uncertainty associated with the result of one bottle. The left-hand side represents the “pure” effect due to between-bottle inhomogeneity, as estimated from an analysis of variance.

A discussion about various approaches to obtain an uncertainty estimate that accounts for insufficient repeatability of the measurement method other than the result of Equation (4) is given in Reference [19]. The influence of the repeatability standard deviation on s_{bb} can be accounted for using

$$u_{\text{bb}} = \sqrt{\frac{MS_{\text{within}}}{n}} \sqrt[4]{\frac{2}{\nu_{MS_{\text{within}}}}} \qquad (6)$$

where MS_{within} is equal to the repeatability variance of the measurements used in the between-bottle homogeneity study.

This expression is based on the consideration that a confidence interval can be developed for s_{bb}, and that the half-width of the 95 % confidence interval, converted to a standard uncertainty, can be taken as a measure of the impact of the repeatability of the method on the estimate of s_{bb} The expression is an example of how the inability of estimating inhomogeneity can be accounted for. Alternatives can be developed but they should meet the criteria as given in Equation (5).

7.10 Within-bottle homogeneity

Within-bottle homogeneity is an issue that only arises when the bottles (units) of the candidate RM can be subsampled. In many cases, it is not possible to obtain an exact estimate of the variance due to within-bottle heterogeneity. The repeatability of the test method will to some extent always be contained in the estimate for within-bottle homogeneity. This makes the estimate for s_{wb} always “safe”, i.e. greater than the actual uncertainty. Figure 3 shows the layout of a within-bottle homogeneity study.

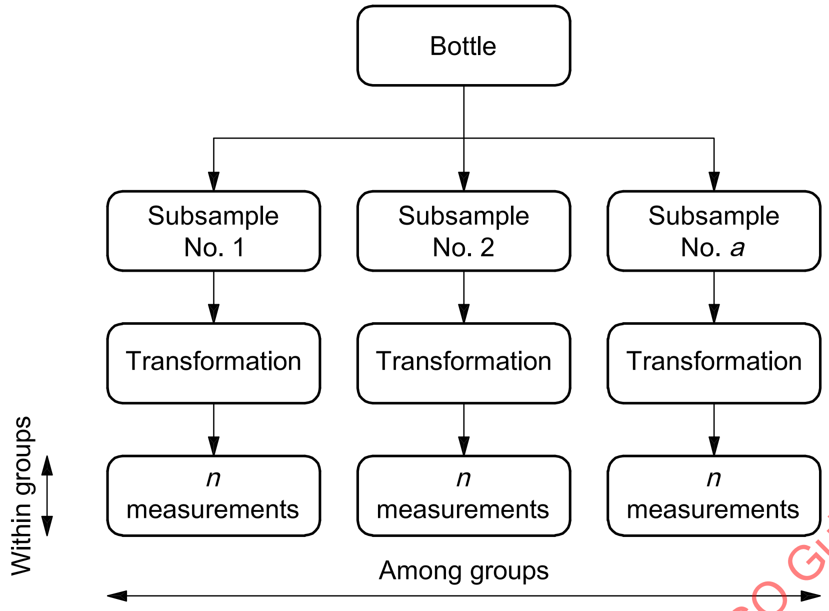


Figure 3 — Layout of a within-bottle homogeneity study
(from Reference [22])

Multiple test portions are drawn from a sample, which can usually only be transformed once (Figure 3). There are notable exceptions (e.g. the use of X-ray fluorescence) where multiple measurements on the same test portion are possible. In these cases, a one-way ANOVA approach may be considered, just as in the case of a between-bottle homogeneity (see A.1). The relevant standard deviation is the standard deviation between groups, where a group represents a subsample.

The minimum sample intake is determined by carrying out a within-bottle homogeneity study for different test portions. As the within-bottle homogeneity standard deviation depends on the number of particles carrying a certain property, it is possible to determine the minimum number of particles (or minimum test portion). This minimum is the smallest sample intake for which the standard deviation of the test portion equals the repeatability standard deviation of the measurement method.

The minimum sample intake may be determined experimentally or through extrapolation. Extrapolation of the within-bottle standard deviations obtained from different subsample sizes may be used to find the smallest subsample size that does not affect the repeatability of measurement for that particular parameter. Due to the fact that the within-bottle homogeneity standard deviation is usually an overestimation, the minimum sample intake will usually be an overestimation too.

Another approach to the problem is to demonstrate for a particular sample intake that the standard deviation over the test portions equals the repeatability standard deviation of the measurement method. Such an assessment can be done by comparing the variances with a χ^2-test (see ISO Guide 33 for details). The sample intake used in such an experiment may be set as the minimum sample intake.

8 Stability study

8.1 Types of (in)stability

There are two types of (in)stability to be considered in the certification of reference materials:

- the long-term stability of the material (e.g. shelf life), and
- the short-term stability (e.g. stability of the material under “transport conditions”).

The long-term stability of a reference material is associated with the behaviour of the RM on the shelves of the producer. The short-term stability is associated with any *extra* effects due to transport of the samples. In some cases it is not possible to maintain appropriate conditions with respect to the stability of the RM during transport and, in this case, allowance should be made for some extra uncertainty in the property values.

For the validity of the uncertainty stated on the certificate of a CRM, a correct estimation of the effects due to both long-term stability and short-term stability are as important as the correct assessment of the batch inhomogeneity (see Clause 7). During the lifetime of the CRM, the validity of the uncertainty on the certificate should be demonstrable, so that the CRM can meet the requirements with regard to stability.

It is often equally important to know what might happen to the sample if proper transport conditions are not maintained. In many cases, a simple verification of the CRM prior to first use may be sufficient to reconfirm the validity of the certificate, whereas in other cases it is evident that the CRM has become useless. This knowledge allows the producer to give better advice and, from the perspective of the user, a better product. Stability studies are therefore not only conducted to assess the uncertainty of measurement associated with the stability of the material, but also to be able to specify proper storage and transport conditions (see also 5.9).

8.2 Designs of experiments

There are two basic experimental layouts for stability studies [13], [23]

- the classical stability study, and
- the isochronous stability study.

In the classical stability study, individual samples prepared at the same time (i.e. as a batch), under identical conditions, are measured as time elapses. In this case, the work is carried out under (within-laboratory) reproducibility conditions, which leads to a relatively high uncertainty, as the instability of the measurement system is also included.

The isochronous stability study has been introduced to allow all measurements of the stability study to take place under repeatability conditions [13], i.e. in one run with one calibration. The word “isochronous” emphasizes that the measurements are all taking place at the same time, rather than distributed over the time span of the stability study, as is the case in the classical approach.

The isochronous approach reduces the scattering of the points over time, thus improving the “resolution” of the stability study. As a consequence, the isochronous stability study will usually lead to a smaller uncertainty than the classical one, depending on the difference between the repeatability and the (within-laboratory) reproducibility of the measurements. A prerequisite for this layout is that conditions can be defined under which degradation does not occur, or at least occurs at a different rate from the conditions selected for storage. The isochronous layout is specifically designed for batch certifications, as it cannot be used in the case where a single artefact is certified.

Both experimental layouts are suited for long-term and short-term stability studies. For a short-term stability study, the behaviour of the material and its property values is studied under (as a minimum) the recommended (specified) conditions for packaging and transport. The more restrictive are the conditions for transport, the smaller the short-term stability study may be kept. It is recommended to apply conditions for transport for which the instability of material is not greater than in the long-term stability study, so that no uncertainty contribution for short-term (in)stability needs to be included in the certification. For several kinds of reference materials (e.g. clinical, biological and environmental reference materials), it is not always obvious that transport conditions can be maintained which allow the effect of transport on the material to be ignored. When no previous experience is available concerning a particular matrix/property combination, a short-term stability study might be carried out at different temperatures, to gain information concerning the appropriate storage conditions, and the necessity to take precautions during transport.

Such a study typically takes about 2 months, but may be extended to 6 to 12 months to obtain additional information about the long-term stability. The range of temperatures for such a continuation may be reduced as appropriate, as the study after 2 months concerns only the storage conditions. Any transport of a CRM can and should be organized in such a way that the time needed for transport is as short as possible. Experience has shown that 2 weeks are feasible, but exceptions may exist. In any case, the short-term stability should include temperatures that might occur during transport (e.g. up to 70 °C and down to −50 °C) for a period that is at least as long as that allowed for transport of the CRM. If such as period is restricted to, for example, 3 weeks, a short-term stability study of 3 to 4 weeks will suffice.

8.3 Evaluation of results

8.3.1 Trend analysis

The first step in the evaluation of data from a stability study is a check of whether any trend in the data can be observed. For small instability problems where the underlying kinetic mechanism is unknown, a linear approximation is a suitable model. In cases where a well-defined mechanism is the reason for the instability, such a model is to be preferred over the (empirical) linear model. The mathematics are somewhat more complex for models other than the straight line, but the evaluation runs in the same fashion, using the F-test to test the trend for significance.

In the absence of a well-defined kinetic mechanism, the basic model for a stability study can be expressed [24] as

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{7}$$

where β_0 and β_1 are the regression coefficients, and ε denotes the random error component.

EXAMPLE An RM containing a radioactive isotope is an example of a parameter with a well-defined kinetic mechanism, in this case a radioactive decay.

The random error component, ε, may be composed of only random error, but it may also contain one or more systematic factors. In the case of stability studies, X denotes time, and Y the property value of the candidate CRM. For a stable reference material, β_1 is expected to be zero. The development of expressions for estimates for the parameters β_0 and β_1, as well as the computation of variances of different kinds, follows the same paths as the development of the expressions for analysis of variance, as shown in Reference [20].

Given a set of n pair-wise observations of Y versus X, for each Y_i the following expression can be developed

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{8}$$

Often, more than one value of Y_i will be available for each X_i, due to repetition of measurement, the use of more than one bottle per point in time, etc. These aspects should be included in the model of the particular stability study. For the trend analysis however, the average result of all bottles at time X_i can be used. Based on this clause and Reference [20], these extensions can be developed quite straightforwardly.

The regression parameters can be computed from the following expressions. For the estimator for the slope, the following expression can be used:

$$b_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \tag{9}$$

The estimate for the intercept can be computed from

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{10}$$

From error analysis, the expressions for the standard deviations in b_1 and b_0 can be computed. The estimated standard deviation of b_1 is given by

$$s\left(b_1\right) = \frac{s}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}} \tag{11}$$

where

$$s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - b_0 - b_1 X_i\right)^2}{n-2} \tag{12}$$

The estimated variance of b_0 is given by

$$V\left(b_0\right) = V\left(\bar{Y} - b_1\bar{X}\right) = s^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right] = \frac{s^2\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \tag{13}$$

where it should be noted that b_1 and $\bar{Y}$ are uncorrelated [24].

Based on the standard deviation of b_1, a judgement can be made. Using Equation (11) and an appropriate t-factor (the relevant number of degrees of freedom equals $n - 2$), b_1 can be tested for significance. Although this method is quite uncomplicated, it requires the computation of $s(b_1)$, a parameter that is often not calculated by software. Most software does, however, compute an F-table, which can also be used for evaluating the significance of regression (see Table 1).

Table 1 — Analysis of variance table for linear regression

Source of variation	Degrees of freedom	Sum of squares SS	Mean square MS	F
Due to regression	1	$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	MS_{reg}	$F = \frac{MS_{\text{reg}}}{s^2}$
About regression (residual)	$n - 2$	$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	$s^2 = \frac{SS}{n-2}$	
Total, corrected for mean $\bar{Y}$	$n - 1$	$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$		

The mean square due to regression is often denoted as $SS(b_1|b_0)$, to be read as “sum of squares for b_1 after allowance has been made for b_0”. The mean square about regression (s^2) is an estimate for the property denoted by $\sigma^2_{Y.X}$ and called the variance about regression.

The ratio $MS_{\text{reg}}:s^2$ can be tested for significance using the F-tables. Table 1 provides the necessary information with respect to the degrees of freedom. The advantage of using the F-table instead of the method using the t-test is two-fold:

— the F-table is generated by most software systems by default, and

— the F-table can readily be extended to other regression models, which makes it more widely applicable.

Irrespective of what kind of test is used, it should be noted that the outcome is only meaningful if the repeatability standard deviation of measurement, possibly in conjunction with the between-bottle homogeneity, is sufficiently small. It can be demonstrated that if the repeatability standard deviation is comparable to that of the homogeneity study and the characterization of the material (e.g. the determination of the property value), this requirement is met as far as the repeatability of measurement is concerned. The effect of the between-bottle inhomogeneity can be reduced by taking multiple bottles at each point in time. Such an approach can be necessary when s_{bb} is equal to or greater than the repeatability of measurement. If a trend is observed, this usually means that the material cannot be certified. A criterion for such a decision should be based on the (expected) uncertainty associated with the property value of the reference material, the (desired) shelf life, and the trend over this period of time. If the trend is meaningful over the period of the (desired) shelf life with respect to the uncertainty associated with the property value of the RM, then either the material cannot be certified due to lack of stability, or the shelf life should be decreased.

8.3.2 Uncertainty evaluation in absence of trend

A stability study includes the following uncertainty components:

- repeatability of measurement;
- instability of the material;
- instability of the measurement system (in the classical design);
- reproducibility aspects (e.g. operator, equipment), including calibration (in the classical design);
- between-bottle homogeneity (in batch certifications).

From this list, it can be seen that, whenever possible, the isochronous design should be preferred over the classical one, as it reduces the number of components to look at. In a typical isochronous stability study, only three components of uncertainty are left, which can be separated through a fully nested two-way analysis of variance [20], [23]. The uncertainty for a single measurement in such an experiment can be expressed as

$$u^2\left(y_{ijk}\right) = s_{\text{stab}}^2 + s_{\text{bb}}^2 + s_r^2 \tag{14}$$

where s_{stab} is the standard deviation due to instability, s_{bb} denotes the between-bottle standard deviation, and s_r the repeatability standard deviation. The index i runs over the points in time, j over the bottles, and k over the repeated measurements.

As in the case of the homogeneity study, the quality of the estimator s_{stab} depends on s_{bb} (and s_r). Thus, the between-bottle homogeneity affects the quality of the estimator for instability. This is inevitable, as it is a property of analysis of variance [20], [30]. The processing of the results can be carried out by a two-way ANOVA, similar to the case described in A.2. It should be noted that (at least in principle) it is possible to estimate s_{bb} from a stability study [23]. When there are multiple temperatures in a stability study, the s_{bb} estimate obtained for the reference temperature will often be the best one, as for this temperature it is assumed that the material is stable. At temperatures where the material is clearly not stable, the changes in the material might affect the obtained estimate s_{bb}.

The assumption has been made that the homogeneity and stability of the material are independent. This is often the case, but there can be exceptions. When there is considerable between-bottle heterogeneity, it can also be expected that the stability of the material will differ from bottle to bottle, as the stability of the material depends (among others) on the composition of the material. The presence of some destabilizing component, quite heterogeneously through the batch, may also be a reason for such a correlation.

In the classical design, the expression for the uncertainty of a replicate reads as

$$u^2\left(y_{ijk}\right) = s_{\text{stab}}^2 + s_{\text{lor}}^2 + s_{\text{bb}}^2 + s_r^2 \tag{15}$$

where one term has been added, the variance due to lack of repeatability[4], s_{lor}^2. This term represents the stability of the measurement system. The measurements in a classical stability study take place under (within-laboratory) reproducibility conditions, which leads to the complication that the stability of the measurement system cannot be separated from that of the candidate reference material. As a result, the uncertainty for instability will always be greater for the classical design than it will be in the isochronous case.

Another option is to estimate uncertainty of stability via the uncertainty associated with the regression line with a slope of zero [16]. This approach gives a “safe” estimate of possible degradation of the material.

The various budgets (s_{stab}, s_{bb}, s_r) can be determined from the respective mean squares (M). In A.2 the two-way ANOVA is discussed. For further details on using fully nested ANOVA designs, see, for example, References [20] and [21].

8.4 Stability monitoring

8.4.1 Experimental

Monitoring should be envisaged during the lifetime of the CRM. A fundamental problem of stability studies is that theoretically they only account for the past, not for the present or future. Some kinds of degradation or other instability problems proceed very slowly and gradually, but in many cases some abrupt change in properties take place at some time, practically ending the lifetime of the CRM. As these mechanisms are highly unpredictable, it is necessary to monitor the stability.

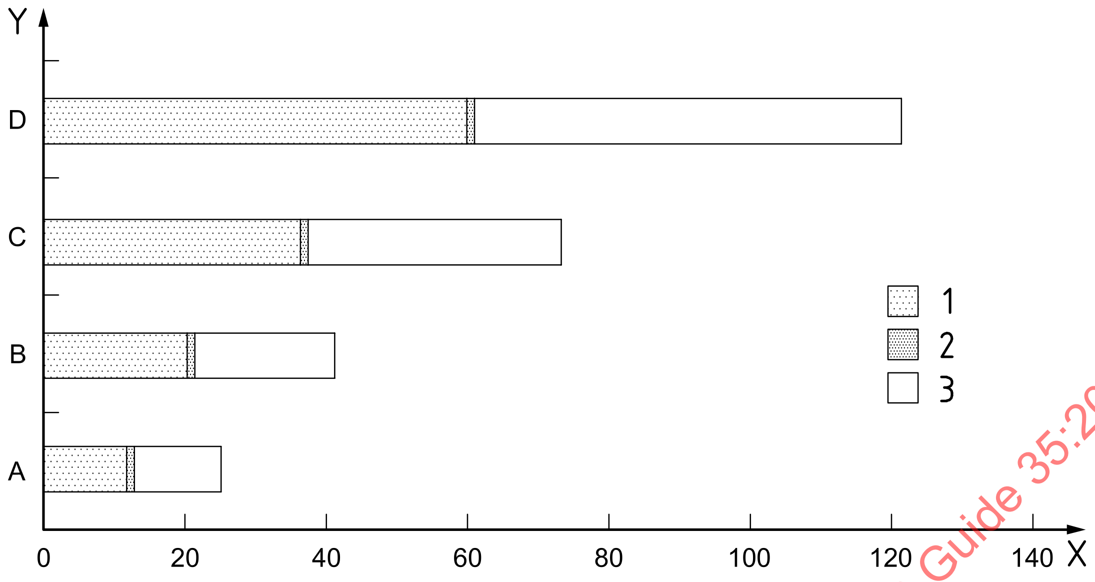
Monitoring usually takes place using the classical design. This is because the isochronous design only provides data at the end of the stability study, whereas for monitoring it is essential that information become available during the lifetime of the CRM. This has no further consequences for the uncertainty associated with the property value of the CRM, in contrast to the other two stability studies, since monitoring only involves the demonstration that the uncertainty on the certificate is still valid. This should obviously be carried out with care, so that not too much uncertainty is added during the verification of the CRM, but there is no necessity to account for these results in the combined standard uncertainty associated with the property value of the CRM.

An important alternative to classical monitoring is to use the isochronous experiment design to carry out a type of semi-continuous stability study. An example of this kind of stability monitoring is shown in Figure 4.

In the logistics phase, the samples are stored at the relevant temperatures. After this phase, all samples should be measured under (ideally) repeatability conditions. Sometimes the measurements take more than 1 day, which means that working under strict repeatability conditions is not possible. The shelf life is determined from the data (see 8.5). It is important to note that the measurements in a subsequent study should take place before the shelf life has ended. Using this design, there is little need to combine results from different stability studies, as the estimators will only slightly improve.

The reason that this type of stability testing cannot be continuous is because of the use of isochronous measurements; since these are carried out in a single run after the period of stability testing, it is necessary to make a “cut” in the stability testing. After such a cut, the uncertainty associated with the property value of the CRM may be reviewed as well, since these new stability data may be used as a renewed estimation of the uncertainty due to instability. In essence, the “cut” is due to the fact that each isochronous stability study is done with a single calibration. Such a calibration is necessary for each new study, whereas in the classical design every data point usually requires a new calibration, giving rise to a “lack of repeatability”.

[4] “Lack of repeatability” means that there exists a reproducibility effect between the points in time in addition to the repeatability of measurement.



Key

X time, months

Y study

1 logistics

2 measurement

3 shelf life

Figure 4 — Semi-continuous stability testing (from Reference [23])

8.4.2 Uncertainty evaluation

The uncertainty evaluation of stability monitoring is quite different from long-term and short-term stability studies. First, it should be noted that stability monitoring does not affect the uncertainty statement of the reference material on the certificate, u_{CRM}. This is logistically impossible and is unnecessary, as will be demonstrated. Ideally, the uncertainty associated with the measurement (u_{meas}) is substantially smaller than u_{CRM}, but this situation may not always be feasible. Furthermore, the measurements should be carried out in such a way that their validity is not demonstrated by using the CRM. One cannot check two things at the same time with one experiment. The validity of the CRM is to be reconfirmed, which can only be valid if the measurement is demonstrably reliable.

If the condition

$$\left|x_{\text{CRM}} - x_{\text{meas}}\right| \le k\sqrt{u_{\text{CRM}}^2 + u_{\text{meas}}^2} \qquad (16)$$

is met, where x_{CRM} denotes the property value of the CRM, x_{meas} the observed value during measurement, and k is an appropriate coverage factor at a level of confidence of 95 %, then the material may be considered to be sufficiently stable, and the stability is demonstrated (provided that the method of measurement is unbiased).

If these experimental conditions are fulfilled, both the property value and its expanded uncertainty are reconfirmed. Under these conditions there is no need to increase u_{CRM}, as the uncertainty from measurement is something which must be accounted for separately. This is true both for monitoring and for the normal use of the CRM. It should, however, be noted that for the sake of the validity of the monitoring measurement, u_{meas} should be as small as possible, and should certainly not exceed u_{meas} from a typical user of the CRM, who will use a similar approach for verifying his measurements.

If the monitoring or continued isochronous stability testing indicates that the property value is no longer valid within its uncertainty, there are two legitimate options (see also 6.7):

- withdrawal of the (certified) reference material, or
- recertification.

8.5 Determination of the shelf life in relation to the long-term stability

The shelf life can be related to the standard uncertainty due to long-term stability, as follows. The basis is the absence of a significant trend in the stability data. Given

$$Y(b_0, b', X) = Y_0(1 + b'X) \tag{17}$$

where it is assumed that the property value Y decreases linearly from the initial value Y_0 with a constant relative degradation rate b' as a function of time X. The uncertainty associated with the property value of the CRM can be estimated by propagating the uncertainties $u(Y_0)$, $u(X)$ and $u(b')$ of the dependent variable Y to the independent variables Y_0, X and b'

$$u^2(Y) = \left(\frac{\partial Y}{\partial Y_0}\right)^2 \cdot u^2(Y_0) + \left(\frac{\partial Y}{\partial X}\right)^2 \cdot u^2(X) + \left(\frac{\partial Y}{\partial b'}\right)^2 \cdot u^2(b') \tag{18}$$

As shown in Reference [25], the last term may be taken as a basis to relate the standard uncertainty due to long-term stability to the shelf life. It should be noted that the partial derivative equals X, the time elapsed since the certification. Using a first-order approach, at a time X

$$u_{\text{lts}} = Y_0 X u_b \tag{19}$$

This expression forms a basis to make allowance for the uncertainty due to long-term instability of the samples, in absence of significant degradation for a given shelf life X.

9 Determination of the property values

9.1 General

There are a number of technically valid approaches to the assignment of property values. These include measurement with one or more methods involving one or more laboratories. An appropriate approach can be chosen depending on the type of reference material, its end-use requirements, the qualifications of the laboratories involved, the quality of the method or methods, and the ability to estimate the uncertainty associated with the characteristics realistically.

Both this clause and Clause 10 are limited to the case where single property values are determined. In several areas, the properties of a CRM of interest may take up other forms, such as spectra. In principle, the contents of Clauses 9 and 10 also apply to these cases, but the mathematics are more complex than for property values. It requires a good knowledge of statistical modelling techniques to apply the concepts as given in this Guide. The aspects of between-bottle homogeneity, long-term stability and, if applicable, short-term stability are also applicable to the cases where properties such as spectra or curves are concerned.

9.2 Establishing and demonstrating traceability

In its role as a measurement standard, the property values of a CRM need to be traceable to appropriate units and/or references. There are several possibilities to achieve this; the appropriate choice should be made based on the intended use of the CRM. The following models exist:

a) if possible, property values should be made traceable to SI units, and expressed in the corresponding units;

b) many RMs form an (accurate) realization of a unit defined by a standard method; these RMs should be made traceable to a result obtained by strictly following this standard method and/or a standard operating procedure developed on the basis of the standard method;

c) RMs can be made traceable to other measurement standards or artefacts, including CRMs and RMs.

There are also conventional scales that are maintained at least partly through reference materials including, for example, the scales for pH and octane number of gasoline. For the pH measurement, the internationally agreed primary realization is through a Harned cell [26]. Often, but not always, conventional scales are maintained by following a fixed recipe to prepare RMs to establish such a scale. When available, such a recipe should be strictly followed.

For many matrix reference materials, the situation is more complex. Although the determination of the property value itself can be made traceable to appropriate units through, for example, calibration of the measurement equipment used, steps like the transformation of the sample from one physical (chemical) state to another cannot. Such transformations may only be compared with a reference (when available), or among themselves. For some transformations, reference methods have been defined and may be used in certification projects to evaluate the uncertainty associated with such a transformation. In other cases, only a comparison among different laboratories using the same method is possible. In this case, certification takes place on the basis of agreement among independent measurement results (see Clause 10).

Traceability of results of measurements is usually assured through calibration against appropriate standards. For many measurement systems suitable for use in certification projects, this may be achieved by calibrating instruments using measurement standards. These measurement standards may include other CRMs or RMs, where it should be noted that since the traceability of CRMs is explicitly stated, their use is preferred over the use of other RMs. The adequacy of the measures taken to ensure proper calibration of equipment and the traceability of results can be verified by, for example, specially designed and prepared control samples (such as a sample otherwise used for calibration). For this specific purpose, these samples are provided without value (and uncertainty), thus allowing assessment of the calibration procedure. CRMs may also be used to demonstrate the validity of the result obtained from a measurement in a measurement campaign.

The transformation of a sample from one physical (chemical, biological) state is often an important part of a measurement method. In some cases, no options are available to verify these steps in the measurement chain. In Figure 5, some typical options for establishing and/or verifying the traceability of a measurement result are summarized [27], [28].

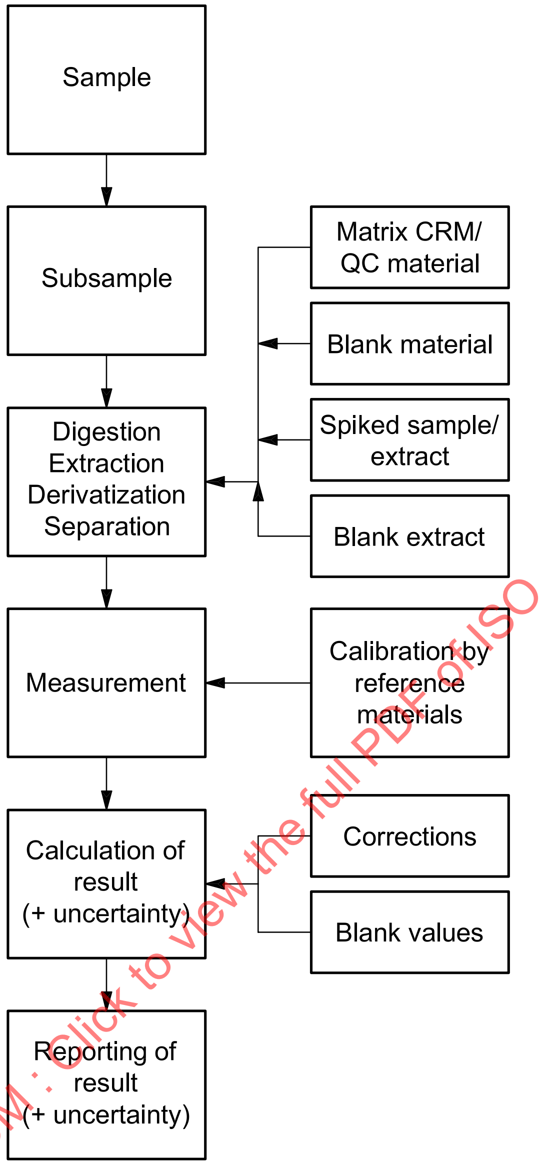


Figure 5 — Part of a measurement chain

Matrix CRMs, matrix RMs and quality control (QC) materials may be used to demonstrate the validity of the measurement result when measured alongside the unknown to be characterized. The use of these materials may support the traceability of the results, but cannot be considered as a direct demonstration of traceability. They demonstrate to a certain extent (see, for example, ISO Guide 33 [9] for details) control over the measurements, which is a prerequisite for achieving traceability of the results used for the certification of the candidate reference material.

Blank matrix materials, blank extracts, etc., may be used to demonstrate that the measurement method provides a result not significantly different from zero when the characteristic of interest is not present (as often done in composition measurements), or to establish a correction or correction factor (+ uncertainty).

Calibration should take place against measurement standards that are traceable to appropriate references. CRMs may be used for this purpose, as long as they are suited for this purpose. The calibration should be appropriate for accurate measurements, thus not introducing any unnecessary extra uncertainty. The reference chosen may be an SI unit (e.g. for composition measurements and many physical quantities), or a conventional scale (e.g. for method-defined characteristics).

Spiked materials, spiked blanks, etc., may be used to validate parts of the measurement chain, or to assist in the process of assigning values to a material. The degree to which this step establishes or demonstrates traceability depends on how these spiked samples are prepared and how their values have been assigned.

Other aspects that may need to be under control for establishing and/or demonstrating traceability of measurement results include

— sample weighing,

— purity of reagents, solvents, "pure materials",

— calibration status of common laboratory equipment and glassware,

— interferences in the measurement signal,

— appropriate and validated statistical/mathematical techniques for doing calculations (e.g. calibration curves, interpolations), and

— contaminations, losses, flaws in the measurement process.

These are all aspects that can be brought under proper control by validating the method. Any method used in a certification project should be properly validated, and it should be demonstrable that any result obtained with the measurement method meets the specifications established during the validation of the method. A laboratory comparison may be part of a method validation. For conventional methods (i.e. method-dependent properties), an interlaboratory validation of the method is essential; for other methods, it is highly recommended to run an interlaboratory validation.

9.3 Practical approaches

Where technically possible, candidate CRMs are generally characterized on the basis of accuracy. Thus, a certified value generally represents the present best estimate of the "true" value. In some cases, the measurement cannot be understood in terms of a "true value", so that an assigned property value for use with a specified method is adopted. Here the certification of the RM does not require a measurement campaign, 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 should take into account all systematic and random effects inherent in the measurement process, as well as any material variability between samples (homogeneity) and in time (stability).

ISO Guide 34 [10] distinguishes between four basic approaches to the characterization. They are implemented in many different variants by producers and certification bodies of RMs, as follows:

a) measurement by a single (primary) method in a single laboratory;

b) measurement by two or more independent reference methods in one laboratory;

c) measurement by a network of laboratories using one or more methods of demonstrable accuracy;

d) a method-specific approach giving only method-specific assessed property values, using a network of laboratories.

The most important aspect to be considered when choosing an approach for certification is to what extent different components of uncertainty will contribute to the uncertainty associated with the property value. Furthermore, it is important that the property value assigned and its uncertainty are demonstrable. Approach a) for certification is often limited to cases where a calibration of an artefact takes place. For certifying chemical compositions, 9.5.2 gives a typical example where approach a) has proved to be valid. For most matrix CRMs, approach a) is associated with the risk of missing matrix effects and/or interferences. For these CRMs, it is recommended to have at least two independent results from independent (primary) methods, carried out by different groups, in order to minimize such a risk. The choice of the best approach for the characterization of an RM thus depends on both the available methods and the matrix of the RM.

9.4 Measurement design

9.4.1 Measurements by one or more (reference) methods in one laboratory

An important group of measurement methods that can be used in particular for approach a), but certainly also for approach b) is formed by primary methods of measurement. In the field of chemical measurement, such a method is defined [29] by CCQM as follows.

> "A primary method of measurement is a method having the highest metrological properties, whose operation can be completely described and understood, for which a complete uncertainty statement can be written down in terms of SI units.
>
> A primary direct method measures the value of an unknown without reference to a standard of the same quantity.
>
> A primary ratio method measures the value of a ratio of an unknown to a standard of the same quantity; its operation must be completely described by a measurement equation."

From this definition, it follows that one of the strategies of assigning a value to an RM is to use a primary method of measurement [30]. It is, however, not always possible to use a primary method of measurement, since they do not exist for all quantities. For many RMs, the measurement of the quantity of interest is so complex that it cannot be described to an extent necessary for establishing a complete uncertainty statement in terms of SI units.

The CCQM has identified several methods with the potential of being a primary method of measurement [31]:

- isotope dilution with mass spectrometry (IDMS);
- coulometry;
- gravimetry;
- titrimetry;
- determination of freezing point depression.

NOTE It is expected that this list will be extended later.

Gravimetry is extensively used as preparation technique for gas mixtures [32] and solutions. Freezing point depression may be used as a direct method to determine the purity (as an amount of substance fraction) of a material [33]. IDMS is widely used as a method for assigning values to materials and for high-quality measurements for other purposes.

9.4.2 Strategies involving multiple laboratories

9.4.2.1 Concepts

The concept of the determination of the property values of an RM based on agreement among methods and/or laboratories is based on at least two assumptions:

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

b) (by implication) the differences between individual results, both within and between methods/laboratories, are statistical in nature regardless of the causes (i.e. variation in measurement procedures, personnel, equipment, etc.).

Each method/laboratory mean is considered *a priori* an unbiased estimate of the characteristic of the material. Usually, the mean of method/laboratory means is assumed to be the best estimate of that characteristic. Furthermore, each of the results obtained in the collaborative study shall fulfil the requirements with respect to traceability as stated in 9.2. In the case of very irregular distributions of results of the campaign, 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 more appropriate to estimate the property value.

In practice, the size of the method/laboratory population that is available to such a programme is limited. In most cases, therefore, a random-design model cannot be fully implemented. Furthermore, it should be noted that the assumption that all results of the various methods/laboratories belong to the same population should be handled with care. Even at the "state-of-the-art" level, differences in performance characteristics of methods as well as differences in uncertainty of laboratories can exist, which might invalidate the assumption of a single population.

Furthermore, for this kind of approach to be valid, it must be assumed that all results of all methods and/or laboratories involved are made traceable (see 9.2) to appropriate stated references. These references are usually carefully selected when designing the collaborative study to obtain the property value.

Often, the involvement of multiple laboratories is necessary in such a collaborative study in order to randomize errors associated with certain parts of the chain of operations necessary to obtain a value. Such steps typically include the transformation of test portions (e.g. extraction, destruction of the sample), further treatment of the transformed sample (e.g. clean-up), and sometimes also impact of different methods of measurement. In these cases, where it is not feasible to perform a full evaluation of measurement uncertainty of a result obtained from a single method (and within a single laboratory), the approach as described in this clause is to be preferred over the approaches given in Clause 10.

The general procedure for the characterization of a candidate CRM by means of a collaborative study is outlined schematically in Figure 6. Each stage can be treated as being distinct and possesses criteria that must be satisfied before proceeding to the next stage.

9.4.2.2 Management and schedule

The management of the collaborative study is primarily the responsibility of the organization responsible for the certification. It should provide sufficient guidance to all involved to ensure the smooth implementation of the work. To be successful, the collaborative study shall have a well-defined objective, be effectively designed and efficiently organized with clear, concise guidelines with which participating laboratories and/or operators can readily comply. Participation, either as operator or as laboratory, in such a programme implies agreement to adhere to these guidelines. These guidelines consist of the time objective, number of units, number of replicate determinations per unit, measurement methods, test portion size where applicable, etc.

The guidelines to be developed should account for the considerations given in 9.2 and 9.3, and translate them into clear instructions so that all parties involved are aware of the requirements with respect to quality and traceability of the measurement results. The guidelines should also contain mechanisms for verifying assumptions being made about the data and the quality thereof.

The organizer shall set the time schedule (i.e. the dates when the samples are to be distributed, the mode of dispatch) and provide clear instructions to dispatchers and receivers with respect to how the samples should be stored and treated. The organizer should also indicate when the measurement results are to be reported. This schedule shall be agreed upon between the organizer and the operators/laboratories involved.

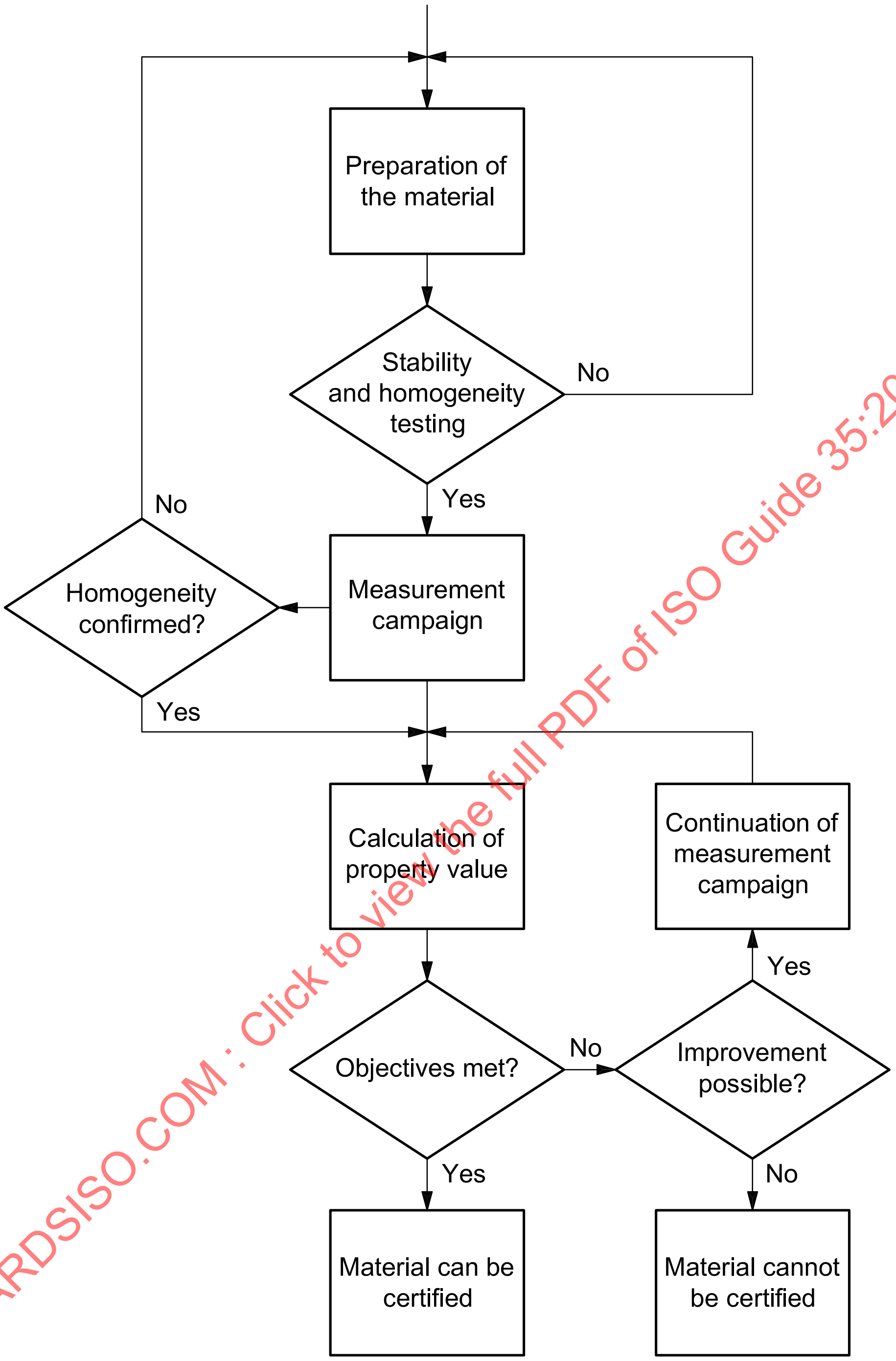

Preparation of the material
Stability and homogeneity testing
No
Yes
Measurement campaign
Homogeneity confirmed?
No
Yes
Calculation of property value
Continuation of measurement campaign
Yes
Objectives met?
No
Improvement possible?
Yes
No
Material can be certified
Material cannot be certified

Figure 6 — Layout of collaborative study

9.4.2.3 Technical requirements

9.4.2.3.1 Required number of results

Frequently, the number of measurement methods available for the determination of a specific characteristic is very limited. When possible, the results from different methods should be checked in order to see whether they agree within their respective uncertainties. If this is the case, then a mean value may be built based on this agreement among methods.

When an evaluation of measurement uncertainty of an RM is to be established through randomization of as many factors as possible during the campaign, then a multilaboratory approach is to be preferred. These ‘laboratories’ may also be different departments or groups within one institute. The minimum number of participating laboratories in the campaign for the characterization of an RM varies with the complexity of the necessary measurement procedure(s). The more complex the procedure, the larger the between-laboratory variation can be expected to be, thereby necessitating an increasing number of participating laboratories to achieve a property value having a satisfactory uncertainty. In practice, unfortunately, the more complex the procedure, the fewer are the groups/laboratories capable of performing it. In extreme cases, the certifying organization may be forced to forego an interlaboratory programme altogether for certain specialized candidate RMs.

If well-established methods of measurement exist for the properties of interest, then the number of laboratories/groups involved in the characterization can be as small as two or three. A typical example of such a case is the use of primary methods of measurement on matrix RMs. When the statistical and metrological control is less, but still adequate to accept (in principle) every technically valid result, the minimum number of laboratories involved is typically six to eight. Finally, if there is a non-negligible chance of having statistically as well as technically invalid results (i.e. limited statistical control), the minimum number of laboratories should be at least 10 and preferably 15. This minimum number allows data to be scrutinized with the aid of outlier treatment techniques, and allows the achievement of an adequate level of uncertainty for the property values thus established.

A further variable to be taken into consideration is the number of methods available and a balanced representation of these methods should be included in the collaborative study. For the case where no primary methods are available, ideally about three methods (when available) should be used by six competent laboratories/groups.

9.4.2.3.2 Number of units and replicate determinations

Usually, two units are sufficient, with about six replicates distributed over (at least) 2 days. All replicates are preferably carried out with independent calibrations. If the results of the collaborative study are to serve as a final confirmation of the homogeneity of an RM, however, values of the characteristic for a minimum of three or four units of the RM should be determined by each participating laboratory in order to have a sufficient number of degrees of freedom for the between-bottle standard deviation to be estimated.

9.4.2.3.3 Measurement methods

The organizer of a collaborative study may specify the use of a single method to participating laboratories when a well-established “standard” measurement procedure is available. This approach is also valid for properties that are defined by a specific method (e.g. leaching properties). Otherwise, the organizer should allow each participating laboratory to propose the method of its choice, provided that he has evidence of the validity of such methods. The organizer should strive for good representation of the major methods suitable for the determination of the particular characteristic, and seek agreement among all parties involved about what methods will be used by each laboratory/group.

Furthermore, the organizer should require that all methods used in the campaign are properly validated, that is, the results of each measurement of the campaign can be verified against pre-determined performance criteria. An essential part of any method validation study is to verify its traceability to internationally accepted standards, to an extent relevant for the particular type of measurement. In many new fields of measurement, the method(s) available sometimes undergo only a validation by means of an interlaboratory study as described in the various parts of ISO 5725 [1] to [6], which may be sufficient for method-defined characteristics.

Finally, a meeting with the laboratories/groups involved (prior to distributing the samples and performing the measurements) may help all parties involved to align all actions to be carried out during the collaborative study, and to discuss possible problems and/or pitfalls.

9.4.2.3.4 Reporting results

Two reporting methods may be applied, depending on whether or not the uncertainties of results are to be reported by each laboratory.

If the participating groups/laboratories are required to state their measurement uncertainty, then the measurement result, its expanded uncertainty and its coverage factor are sufficient in principle. Preferably, however, each laboratory should report the complete uncertainty model with all uncertainties, which will facilitate evaluation of any covariances [34], [38] between results.

If no uncertainty information is requested, then the participating groups/laboratories should report individual results (not the average). The number of significant figures reported should comply with the guidelines for the programme.

In both cases, 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 (e.g. in chemical analysis, the decomposition of the sample and separation of the analytes of interest). References to the literature (where appropriate) should be stated.

9.5 Property-related considerations

9.5.1 RMs for physical properties

Traditionally, the most accurate measurements are carried out for fundamental units, their most common multiples and their sub-multiples, in national metrology institutes (NMIs). Here, all sources of errors and uncertainties are investigated in great detail; methods of measurement, often calibration methods, have been improved over many years to reduce and estimate uncertainties. The accuracy of these measurements is usually well established, especially when they have been the subject of (key) comparisons. Reservations should be made for measurements where there have been no comparisons, such as cases where laboratories use in-house validated methods. Demonstrating the performance of a method in comparison with another laboratory is one of the cornerstones of assuring quality and traceability of measurement results, irrespective of a laboratory's designation. Thus, any new laboratory being established needs extensive comparisons to ensure that its own estimates of values and their uncertainties are equivalent and that no important factors contributing to the uncertainty have escaped its attention.

Special attention should be paid to physical properties that cannot be determined under a calibration regime. Usually, the uncertainty evaluation of results from the methods and test results as obtained in a certification study are not as well established as under a calibration regime. This aspect should be considered when characterizing materials for properties such as thermal conductivity, impact roughness, creep or compressive strength. In these cases, a collaborative study (see Clause 10) may be a more appropriate approach for a characterization. A further complication is that for many of these tests, no (key) comparisons are organized. In these respects, there is no fundamental difference between the certification of physical properties and, for example, chemical composition.

Key comparisons, as well as other types of laboratory comparisons, add confidence to the uncertainty computed by the metrology laboratories individually. Ideally, the results of such comparisons should be used to improve uncertainty models and/or estimates of their variables and/or uncertainties. Even when no improvements of these kinds are necessary, participation in these comparisons is an important means of demonstrating a (metrology) laboratory’s measurement capability. Comparisons allow detection of errors that were not properly taken into account, and situations where some parameters influencing the measurements are not sufficiently well controlled and/or estimated.

Characterizing a reference material on the basis of results of a single (metrology) laboratory may, despite all efforts, still imply a risk which should not be overlooked. When the certification of a physical property or quantity is undertaken, it is important to have a (key) comparison between the major metrology laboratories, followed by a full discussion of the results with all participants to resolve any possible discrepancy. If the NMIs are not themselves involved in the measurement, complete traceability of the participating laboratories to the respective national laboratories should be established before starting.

If more than one method is possible, and if these methods appear equally valid, it is important to compare them. They should, however, be of the same level of accuracy as otherwise, for the purpose of certifying a candidate reference material, the more accurate method is to be preferred.

At the limit, there can exist situations where a 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.

9.5.2 RMs for chemical composition

9.5.2.1 Purity of CRMs

Pure substances form the basis for many traceability chains in chemistry. The adjective “pure” refers to an idealized situation: no substance is 100 % pure, there will always be small impurities. The certification of substances for purity is an essential cornerstone of traceability in chemical measurement. The CRMs are either used by laboratories to prepare calibration standards, or they are used to certify or prepare other CRMs, such as, for example, solutions or gas mixtures. For spiking (see 5.7.4), it is essential that the materials used have been thoroughly characterized for purity. Furthermore, to enable the conversion from mass fractions into amount-of-substance fractions, the full table of impurities and their mass fractions should be stated.

Apart from the direct method through, for example, calorimetry (freezing point depression), the purity is often determined by difference, involving analytical chemical techniques, as follows:

- a list of possible impurities is compiled, often based on the manufacturing process that was used to produce the substance;
- each of the possible impurities is determined in the substance to be certified;
- the purity is computed by difference.

The measurements necessary to determine the impurities are often difficult, since most impurities will be close to the detection and/or determination limits. The measurement of impurities may involve multiple methods/laboratories, including approaches as outlined in 9.4.2. This may well result in high relative uncertainties for the amount-of-substance fractions of these impurities. Also, the evaluation of the uncertainties is far from uncomplicated, as the vicinity of mathematical limits (amount of substance and mass fractions are only defined between 0 and 1) may create additional problems, including negative estimates for such fractions (see, for example, appendix F of Reference [15]).

The model for the amount of substance fraction of the main component y as a function of k impurities with amount of substance fractions x_i is

$$y = 1 - \sum_{i=1}^{k} x_i \tag{20}$$

Assuming independence among the amount of substance fractions of the impurities (which is often the case), the combined standard uncertainty associated with the amount-of-substance fraction of the main component is

$$u^2(y) = \sum_{i=1}^{k} u^2(x_i) \tag{21}$$

which follows directly from applying the uncertainty propagation formula from the GUM to the model [5]. It frequently happens that some of the amount of substance fractions x_i are zero, due to the fact that either these impurities are truly absent, or that their levels are below the detection limit of the measurement method. If the detection limit is used to establish the value for an impurity, this limit should also be used to establish a standard uncertainty, appreciating that this limit determines the highest level of the particular impurity that cannot be detected.

[5] This expression is exact as the model is linear and the xi values are assumed to be independent.

9.5.2.2 Synthetic RMs and gas mixtures

Synthetic reference materials, such as solutions and gas mixtures, are widely used for calibration. These CRMs are often made by means of gravimetry. When gravimetry is used to prepare a bulk solution which is then subjected to a subsampling and bottling procedure, the batch may be certified as follows:

- step 1: the gravimetric composition is taken as basis for the certification;
- step 2: the gravimetric composition is verified using a suitable analytical method;
- step 3: a homogeneity study is carried out to determine the between-bottle variability;
- step 4: a stability study is carried out to determine the long-term stability.

The effects in terms of measurement uncertainty from steps 2 and 3 are expected to be small (see Clauses 7 and 8 for details on these steps) but should be included. If they are negligible, the magnitudes of these effects are such that they do not impact the standard uncertainty associated with the property value of the CRM. The verification uncertainty is (to an extent depending on the ability to verify the composition) included in the model, together with the gravimetric uncertainty. The combined standard uncertainty associated with the property value of the CRM would become

$$u_{\text{CRM}} = \sqrt{u_{\text{grav}}^2 + u_{\text{ver}}^2 + u_{\text{bb}}^2 + u_{\text{lts}}^2} \qquad (22)$$

For gas mixtures for example, models have been established for the gravimetric preparation, which may to some extent also find application outside this particular area. ISO 6142 [35] describes the preparation and value assignment for the gravimetric preparation of gas mixtures. For batches of gas cylinders, however, other techniques are often used [36]. These batches are then certified making use of gravimetrically prepared gas mixtures, used as calibration RMs. A detailed uncertainty model, based on the methods described in ISO 6142 is given in Reference [32]. Based on the preparation, the composition of the mixture can be expressed as amount of substance fractions of the components of the parent gases. These parent gases are the gases used for preparation of the mixture, which can themselves be mixtures or pure gases (see 9.5.2.1).

The model of ISO 6142 accounts for the impurities of the parent gases, which is an important prerequisite for preparing gas mixtures that are to be made traceable to an SI unit, in this case the mole. Some effects may not be incorporated in the model as given, for example that the composition that enters the cylinder is not necessarily the same as the one that can be taken from the cylinder. This may be due to adsorption/desorption effects. Furthermore, a quality control check is required for any mistakes that might have been made during preparation. In order to accomplish this, the gravimetric amount of substance fraction is usually compared with measured (“verified”) amount of substance fraction, and it is assumed that the composition from preparation is not different from that from analytical verification [35], [37].

9.5.2.3 Spiking blanks and blank matrices

The method followed for the preparation of synthetic CRMs may also be used for the spiking of blanks and/or blank matrices. The only extra complication is the verification of whether the material to be spiked is really a “blank”, or that some small amounts of the substances to be spiked can be demonstrated to be present. In that case, these amounts should be taken into consideration in the model for the property value, and therefore also in the uncertainty model.

The approach to certification can be similar to that in 9.5.2.1, but in cases where problems exist in the transformation of the sample (see 9.2), or in determining the uncertainty of this step, one of the approaches as given in may be used as well. The actual choice depends very much on the considerations given in 9.2 and 9.3.

9.5.3 Characterization of conventional properties

In chemistry, biochemistry and other areas of measurement, 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 how 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.

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 inspection is to measure or test a sample of known properties. If the results are satisfactory, it means that the machine is in good condition and that therefore the results may be considered traceable to the measurement scale established by the relevant written standard.

Of course, the characterization work to establish CRMs for such properties or measurement scales requires 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 carried out 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.

10 Data and uncertainty evaluation

10.1 Models

Even the most empirical data assessment follows a single or a certain set of rules, which can not necessarily be formulated in (mathematical) equations. For the technical purposes considered here, rules and relationships can be expressed in the form of mathematical equations, and are called models. Two basic forms must be dealt with.

a) The first is theoretical models describing established (most often physical) relationships between the influential variables of a measurement procedure and the property value under determination. Equations are considered as exact and used for value computation. In a second step, the same measurement model is used for propagating uncertainty estimates attributed to the influential variables.

b) The second is empirical models describing assumed relationships between random variables and certain parameters of the underlying (assumed) distribution(s). They are used for developing procedures for the determination of reasonable statistical parameter estimates of the random variables involved.

There also exist models that are a mixed form of the two, often called semi-empirical models. Examples of strictly empirical models can be found in Annex A, where ANOVA procedures for different numbers of influential factors are developed. Theoretical and semi-empirical models and their use in uncertainty estimation are described in the GUM and in Reference [15].

10.2 Data formats

Except for approach a) in 9.3, where the data format consists of one or more single measurement results of one method and a sensible uncertainty statement (see below), reported data can be formatted in matrices of data containing

1) an appropriate estimate for the property under determination (mean or average) and an uncertainty statement, or

2) a certain number of single measurement results of the property under consideration (replicates),

for each participating laboratory. Format 1) enables uncertainty-based evaluation (see 10.7) assuming that participating laboratories have suitable measurement models, while format 2) (the more “classical” one) requires statistical evaluation based on assumptions with respect to underlying distribution functions.

The results submitted by the participating laboratories should be evaluated in accordance with the procedure outlined in Figure 7.

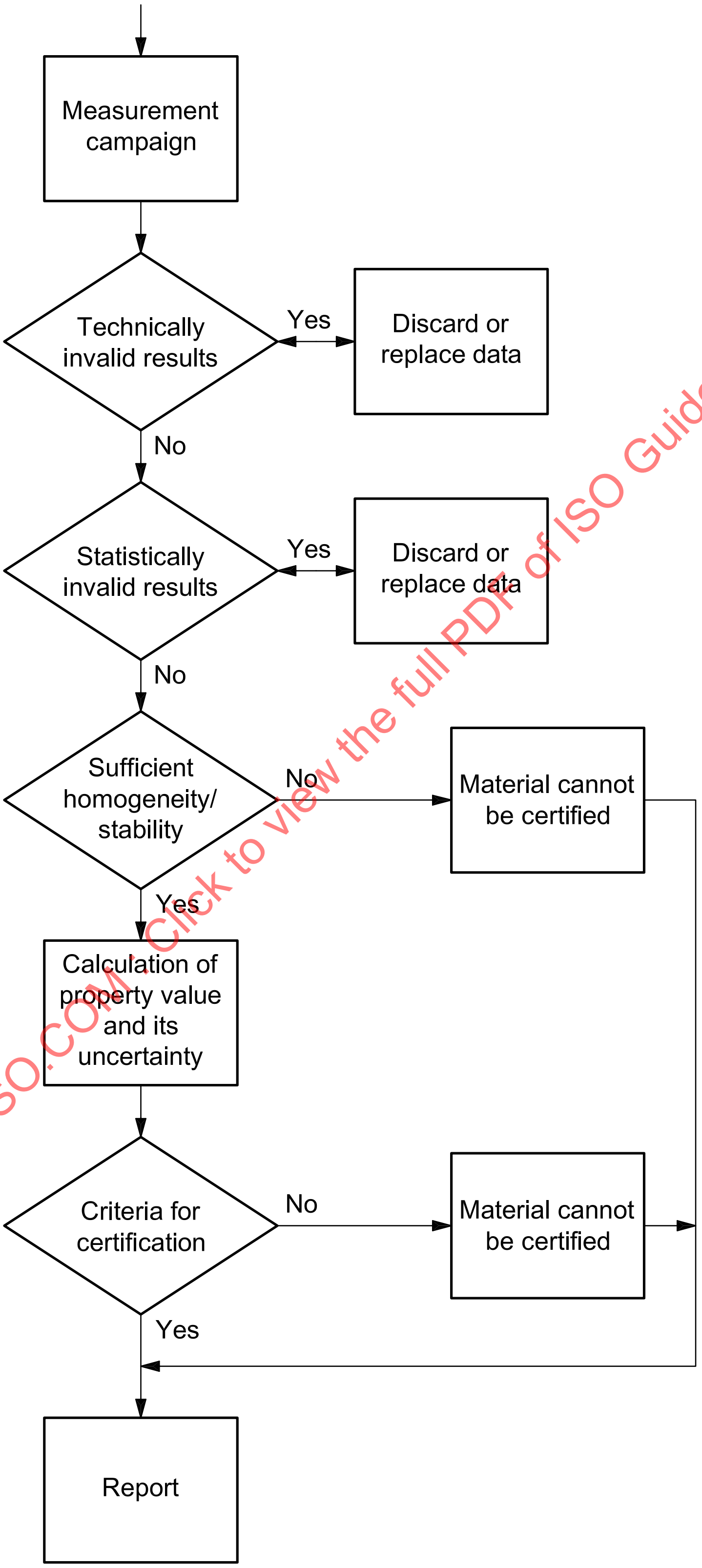

Measurement campaign
Technically invalid results
Yes
Discard or replace data
No
Statistically invalid results
Yes
Discard or replace data
No
Sufficient homogeneity/ stability
No
Material cannot be certified
Yes
Calculation of property value and its uncertainty
Criteria for certification
No
Material cannot be certified
Yes
Report

Figure 7 — Data treatment

For the convenience of processing and for future reference, the results from this type of collaborative study should be grouped on the basis of the characteristic and tabulated systematically. This table should include identification of the group/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.

Two reporting methods may be applied, depending on whether or not the uncertainties of results are to be reported by each laboratory.

If the participating groups/laboratories are required to state their measurement uncertainty, then the measurement result, its expanded uncertainty and its coverage factor are, in principle, sufficient. Preferably, however, each laboratory should report the complete uncertainty model with all uncertainties, which will facilitate evaluation of any covariances [34], [38] between results.

If no uncertainty information is requested, then the participating groups/laboratories should report individual results (not the average). The number of significant figures reported should comply with the guidelines for the programme.

In both cases, 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 (e.g. in chemical analysis the decomposition of the sample and separation of the analytes of interest). References to the literature (where appropriate) should be stated.

10.3 Distributions

Finding appropriate estimators for the expectation of a random variable is closely linked to the (either assumed or determined) underlying distribution, mostly expressed as a probability density function. In many cases, the distribution can be observed graphically by preparing a histogram. If no assumptions are made, data evaluation of a certain number of measured values can entirely be based upon

— an investigation of their distribution using appropriate estimation procedures (e.g. Kernel estimation), or

— the calculation of the relevant estimators following the principles laid down in ISO 3534-1 (calculation of the moments),

provided a statistically relevant number of measured values are available.

Other distributions commonly in use for the quantification of uncertainty contributions are also symmetrical as, for example, the rectangular or triangular distribution (see also GUM).

The approaches given in 10.5 and 10.6 always refer to a normal distribution of data (except for cases where type B evaluations of uncertainty contributions are concerned). For this reason, tests for normality of the available data sets may be necessary to validate the approach selected.

If most of the results form two or more clusters, no agreement among methods/laboratories 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 single property 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.

Slight discrepancies among results from different methods may be overcome by introducing an additional uncertainty component accounting for this aspect. Whether such an approach is of use can only be judged after determining the magnitude of this uncertainty component, and verifying the combined standard uncertainty associated with the property value thus obtained against the criteria set for certification. There are several approaches to this estimation problem. A useful discussion is given in Reference [43].

If most of the results form a single cluster, it may be inferred that the distribution is unimodal. The most obvious choice as an estimator for the characteristic to be determined is the mean of the results.

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

In some cases, the results can be transformed so that they become approximately normally distributed. Some commonly used transformations include logarithmic, square root and exponential forms.

It should be noted that, for the estimation of property values and their uncertainties, the assumption of the distribution function is less critical than for outlier testing. However, as most descriptive statistics are based on the normal distribution, it is recommended to check whether the distribution of the data agrees reasonably with the normal distribution, or that a transformation is possible so that statistical techniques assuming normally distributed data may be used.

10.4 Data screening

Independently of the format, data sets should be inspected visually and graphically before any of the procedures of 10.5 and 10.6 are applied. Any observed anomaly should be scrutinized for possible trivial (transmission error, misprint, etc.) and non-trivial reasons (drop-out, equipment failure, etc.). If errors or failures are confirmed, the corresponding results should be rejected.

Furthermore, the results shall be checked for technically invalid results. Technically invalid results are found by carefully examining the reports of the measurements. A technically invalid result is not necessarily a statistically invalid result. The result in question may fall well within the range of valid results, but it is evident that the conditions under which the result is obtained were not in good order. Technically invalid results should be removed from the data set.

10.5 Data evaluation

10.5.1 Approach A: single method in a single laboratory

When a single value is obtained from the (primary) method of measurement, accompanied by an uncertainty statement, then this value with associated uncertainty is the result of the characterization. If there is a series of values, then usually the mean of such series of values is the result of the characterization. The uncertainty is then usually based on the uncertainty evaluation of the method of measurement, given the value, and should be compared with the standard deviation of the series of values. The standard deviation of the mean and the complete uncertainty budget as established for the method of measurement should be concordant.

10.5.2 Approach C: a network of methods and/or laboratories

The data format can be either

— a series of values, each accompanied by an uncertainty statement, or

— a series of observations of each laboratory.

In the first situation, the uncertainty statements should, as part of the data screening, undergo a plausibility check, for example by checking it with the description of the measurement method. Then the data can be treated as described in 10.7.

When there is only a series of observations, data can be treated as described in Annex A and in References [11] and [12]. A procedure can comprise the following:

a) testing for significant differences between laboratory means as a basis for the decision whether a mean of single values or a mean of laboratory means should be formed;

b) testing for normality of, and outlying values in, the data set chosen according to the decision made under a); outliers can be treated as described in 10.5.5;

c) testing for outlying laboratory variances in the full data set (if necessary, i.e. an abnormal variance is present).

If the dataset is approximately normally distributed, then the mean as chosen under a) is the default choice for the value from the characterization. If the characteristic to be certified is just the mean of means, i.e.

$$\bar{Y} = \frac{1}{p}\sum_{i=1}^{p} Y_i \qquad (23)$$

then the basis for the combined standard uncertainty associated with the mean of means is the standard deviation as obtained from

$$s^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(Y_i - \bar{Y}\right)^2 \qquad (24)$$

The combined standard uncertainty equals then

$$u_{\text{char}} = \frac{s}{\sqrt{p}} \qquad (25)$$

In these formulae, p denotes the number of laboratories. The validity of this expression depends heavily on the independence of the Y_i values. It should therefore be verified to what extent the assumption of independence of the results from methods/laboratories is valid. Especially for large values of p, this verification is critical (see also 10.6.2).

10.5.3 Approach B: Multiple methods in a single laboratory

There are again two possible data formats

— a series of values, each accompanied by an uncertainty statement, or

— a series of observations of each laboratory.

In the first case, each value and accompanying uncertainty statement is scrutinised as described in 10.5.1, for each of the methods involved in the characterization. Then the approach as described in 10.7 can be applied.

The approach of 10.5.2 can be used for the second case. At least for one of the methods, a full uncertainty budget should be available, so that the uncertainty estimated can be verified, and measurement bias can be assessed. (This does not lead to an additional requirement, as all methods involved should meet the traceability requirements outlined in 9.2.)

10.5.4 Approach D: Method-defined parameters

Method-defined parameters may be treated in the same way as described in 10.5.2.

10.5.5 Treatment of outlying values

A single result or an entire set of results is suspected to be a statistically invalid result (an 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 of the detection of outliers depends on the validity of the assumption of the frequency distribution. It is equally essential to have a sufficient number of observations and context information at hand to make proper judgements with respect to outliers. The test for outliers should be the statistician's prerogative, both for selecting valid approaches for outlier testing as well as for carrying out the analysis.

Outlier results can occur at all levels of a collaborative study: single observations, subgroups of observations (e.g. grouped per bottle), or the results from complete methods/laboratories can be observed to be outlying. Outliers should be discarded or, in rare cases (e.g. calculation errors), should be replaced by corrected data. Whenever possible, outliers should be removed only on the basis of the outcome of more than one outlier test. Outliers in variances should only be removed in extreme cases, that is, when they contradict the dataset. Additional measurements are usually not acceptable, as the conditions under which the data were obtained are no longer the same for all results.

Finally, it should be noted that distinction is often made between stragglers and outliers (see for example ISO 5725-2) [2]. As a rule, stragglers should be retained in the data set, whereas outliers should be removed. This will in most cases ensure that the uncertainty estimate does not become an underestimation.

10.6 Uncertainty evaluation

10.6.1 Approach A: single method in a single laboratory

The uncertainty as determined in 10.5.1 is the uncertainty from characterization, u_{char}.

10.6.2 Approach C: a network of methods and/or laboratories

If the values reported by the laboratories are accompanied by uncertainty statements, then the method described in 10.7 is used, and the uncertainty computed is the uncertainty from characterization, u_{char}.

If the laboratories only report a series of values, the standard deviation about the mean is, at least in principle, the uncertainty from characterization, u_{char}. If there are common sources of uncertainty and/or bias, then the standard deviation about the mean is to be complemented with estimates for these sources of uncertainty.

10.6.3 Approach B: Multiple methods in a single laboratory

If the values reported by the laboratories are accompanied by uncertainty statements, then the method described in 10.7 is used, and the uncertainty computed is the uncertainty from characterization, u_{char}.

In the other situation, the standard deviation about the mean is, at least in principle, the uncertainty from characterization, u_{char}. If the check for compatibility with the uncertainty associated with a result of one of the methods used indicates that there is a discrepancy, then allowance should be made for such discrepancy in the uncertainty budget.

10.6.4 Approach D: Method-defined parameters

The uncertainty evaluation of method-defined parameters may be treated in the same way as described in 10.6.2.

10.7 Uncertainty-based evaluation

10.7.1 Basics

When uncertainty information is available, the primary task of the statistician evaluating the results from different methods/laboratories is to combine the results, including their uncertainties, into a single value (the property value) and a combined standard uncertainty. When uncertainties are determined, uncertainty models for the measurements being part of the campaign are available. In order to obtain the best possible result, it is highly recommended that these uncertainty models be collected as part of the evaluation of data.

The property value is usually defined as some kind of mean value, which may be weighted using some predefined weighting scheme, if necessary. When the uncertainty models are available, then the expression for the combined standard uncertainty associated with the property value can be obtained directly using standard propagation of uncertainty formulae (GUM:1993, Clause 5), after adding an additional source of uncertainty that accounts for any significant scatter in the results across laboratories. This is the most straightforward method, but it is not always possible to proceed this way.

A possible alternative is given in Reference [34] and further implemented in Reference [38]. The combined standard uncertainty associated with the property value can be defined by

$$u_{\text{char}} = \sqrt{u_{\text{I}}^2 + u_{\text{II}}^2 + u_{\text{III}}^2 + u_{\text{IV}}^2} \qquad (26)$$

where four types of uncertainty are considered:

- type I: uncertainties specific to a laboratory;
- type II: uncertainties common to all laboratories;
- type III: uncertainties common to groups of laboratories;
- type IV: the disagreement among values from participating laboratories.

Implementing this uncertainty evaluation by hand is laborious and quite sensitive to mistakes. Based on the measurement models, it is unnecessary to do this by hand, provided that the information gathered as part of the collaborative study allows identifying correlated variables in the models. Once all uncertainty components are properly registered in a database, this database can be used to develop all different types of uncertainties. This can be carried out in different ways, either by directly determining the necessary expressions for the terms [34] or for instance by χ^2-fitting [6)] [38], [39].

10.7.2 χ^2-fitting

The χ^2-fitting method runs as follows. The following matrix-equation can be developed:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} a \qquad (27)$$

where a denotes the property value, y_i denotes the results of the measurements as obtained in the collaborative study and $\boldsymbol{x}$ is the design vector. For this particular case, all x_i values are known, as soon as the

[6)] In the literature, this is also known as “least squares” or “least squares fitting”. In principle, χ^2-fitting differs only from “least squares fitting” by a scaling factor, allowing evaluation of the statistic χ^2 after the fitting procedure.

fitting problem is defined. For this situation, $x_1 = x_2 = \dots x_p = 1$. The results $\boldsymbol{y}$ are obviously associated with uncertainties. These will be accounted for in a variance-covariance matrix, defined as

$$\boldsymbol{V}(\boldsymbol{y}) = \begin{pmatrix} u^2(y_1) & u(y_1, y_2) & \dots & u(y_1, y_p) \\ u(y_2, y_1) & u^2(y_2) & \dots & u(y_2, y_p) \\ \dots & \dots & \dots & \dots \\ u(y_p, y_1) & u(y_p, y_2) & \dots & u^2(y_p) \end{pmatrix} \tag{28}$$

It should be noted that this matrix has dimensions p by p, and it is symmetrical (i.e. the upper triangle contains the same elements as the lower triangle). The covariance between each pair of laboratories is evaluated through searching all variables involved in the computation of the result, and identifying the shared ones. For this search, it is not of interest whether these components are only shared by the two laboratories under investigation, or shared by more. The process of pairwise examining any existing covariances means that, during this process, no distinction is made between type II and type III uncertainties (see 10.7.1).

After establishing the covariance matrix $\boldsymbol{V}(\boldsymbol{y})$, the fitting problem may be defined as

$$\varphi(\hat{a}) = (\boldsymbol{y} - \boldsymbol{X}\hat{a})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{a}) \tag{29}$$

The property value can be expressed as

$$\hat{a} = x_{\text{char}} = \boldsymbol{C}\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y} \tag{30}$$

where

$$\boldsymbol{C} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \tag{31}$$

and its variance from

$$V(a) = \boldsymbol{C} = u_{\text{char}}^2 \tag{32}$$

There exist numerous stable algorithms for solving the fitting problem as given in Equation (29). Furthermore, a remark should be made about the matrix $\boldsymbol{V}$. This matrix is equal to $\boldsymbol{V}(\boldsymbol{y})$, provided that the latter contains all uncertainty components relevant to the comparison. If this is not the case, an additional variance-covariance matrix should be added to $\boldsymbol{V}(\boldsymbol{y})$. The construction of such a matrix is analogous to that of $\boldsymbol{V}(\boldsymbol{y})$, whereby often only the diagonal is filled.

The method shown here is a straightforward implementation of the framework as described. It is by no means the only method (see References [40] and [41]). The problem of outliers should be dealt with prior to using this recipe (see also 10.5.5). Likewise, the uncertainty statements should be critically reviewed for their credibility and, if necessary, the weighting scheme should be adapted accordingly [42] (see also 10.5.1). It is well known that χ^2-fitting is sensitive to outlying results as well as problems with weighting.

10.8 Specific issues

10.8.1 Data evaluation using analysis of variance

For strategies C, D, and as part of approach B, analysis of variance (ANOVA) may be used as a tool to process the data. The use of ANOVA can be particularly helpful when assessing uncertainty components such as the between-bottle homogeneity (see, for example, 10.8.2 and A.2), or the between-laboratory standard deviation (see A.3). Otherwise, the mean of means may be computed for these strategies instead.

10.8.2 Confirmation of homogeneity as part of the collaborative study

The results of the collaborative study may serve as a final confirmation of the homogeneity of an RM. For this purpose, a two-way nested design should be followed, in which pq units are used and p laboratories and/or methods are used which each determine the value of the characteristic of q units with n replicate determinations per unit (see Clause 7 for details on a homogeneity study). It is important that this design is strictly followed, to satisfy the requirements for analysis of variance and meeting its underlying assumptions [20]. Clause A.2 provides the necessary statistics to compute the relevant variances.

10.8.3 Combined standard uncertainty of a property value

A whole range of other schemes may be developed to establish a property value based on a series of measurements. Such a property value is typically some kind of mean, or sometimes a robust mean [42]. This subclause briefly discusses how to translate an expression for a mean, including the ones from an ANOVA-based approach, into an expression for the uncertainty of a property value, u_{char}.

The mean of a collaborative study can be defined as follows

$$\bar{x} = \sum w_i x_i \qquad (33)$$

where x_i are the results, and w_i the weightings. No assumption is made concerning the nature of the weightings; they can be based on the number of observations, the uncertainty associated with the laboratory results, or some other scheme. The expression, however, assumes that the weighting scheme is defined in such a way that their sum equals unity.

If from all x_i values uncertainty statements are available, then the uncertainty associated with the property value can be expressed using

$$u^2(\bar{x}) = u_{\text{char}}^2 = \sum w_i^2 u^2(x_i) \qquad (34)$$

provided that all x_i values are mutually independent. This basic method may also be applied to the approaches given in 10.4. When using uncertainty statements, these should be checked for credibility (see also 10.6.1). If no uncertainty data are available, then the calculation of a mean of means (see 10.5.2 and 10.6.2) may be used as an estimate of the uncertainty.

11 Certification

The concept of CRM has been introduced as a special kind of RM. In addition to the characteristics of an RM as defined in ISO Guide 30, a CRM is accompanied by a certificate as described in ISO Guide 31, providing among others the following information:

- the properties of interest;
- their values;
- their uncertainties;
- a statement concerning metrological traceability of the property values.

There are RMs on the market, accompanied by documentation containing the information as described above for CRMs, but the documentation is not called “certificate”, for legal or other (non-technical) reasons. As these RMs must fulfil the same requirements and can be used for the same purposes as a CRM, they are also covered by this Guide. It is understood that these RMs are included in the term “CRM”. This Guide describes the process of preparing a candidate RM so that it may either be certified, or marketed as an RM with a documentation package containing at least the information listed above.

Furthermore, according to ISO Guide 31, the certificate accompanying a CRM is a summary of an extensive programme of work involving selection of material, assessing its suitability, and measuring the properties to be certified. Many users of the CRM will not require any information further to that contained in the certificate but it should be available, either in the form of a certification report (supplied with the CRM or obtainable on request) or otherwise supplied on application to the producer. It is essential to include the name of an officer representing the certifying body indicating that this person accepts responsibility for the contents of the certificate. It is best left to the discretion of the certifying body whether the certificate should also be signed.

Annex A
(informative)

Statistical approaches

A.1 One-way analysis of variance (ANOVA)

Consider the case that there are a groups, and each of them contains n_i members. Ideally, the number of members in the groups should be equal, but in practice this is not always the case. Some data may be "missing", and expressions have been developed to account for these missing data [20], [21] and they are recommended over other methods for treating incomplete data sets. It should be noted that the more incomplete the data set becomes, the poorer the quality of the estimates becomes.

The scattering of data can be expressed in terms of sums of squared differences, also known as "sums of squares". These sums of squares express the scattering at various (hierarchic) levels in the analysis of variance [20]. The so-called mean squares, as obtained from a spreadsheet program, can be converted into variances as follows:

$$s^2_{\text{within}} = MS_{\text{within}} \tag{A.1}$$

$$s_A^2 = \frac{MS_{\text{among}} - MS_{\text{within}}}{n_0} \tag{A.2}$$

where

$$n_0 = \frac{1}{a-1}\left[\sum_{i=1}^{a} n_i - \frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i}\right] \tag{A.3}$$

When there are no missing data, n_0 becomes equal to n. The mechanism as shown allows scatter in the measurements to be attributed to the various uncertainty components influencing the material and measurement process. In the absence of any between-group effect, $s_A{}^2$ is expected to be (close to) zero. If, for experimental reasons, a slightly negative value would be obtained for $s_A{}^2$, then it is set to zero.

EXAMPLE In a between-bottle homogeneity study, s_A is identical with the between-bottle standard deviation s_{bb}. Each bottle is a group.

A.2 Nested random effects in data analysis: Two-way ANOVA

This model may be used when the results of the measurement campaign collaborative study are used to confirm the homogeneity of the material as well as to characterize it. The experimental scheme is illustrated in Figure A.1 for the particular case of an interlaboratory study. When a campaign consists of different methods, the lay-out for the campaign is essentially the same.

The results can be expressed by the equation

$$x_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk} \tag{A.4}$$

where

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

A_i is the error due to method/laboratory i;

B_{ij} is the error due to the jth sample unit within method/laboratory i;

ε_{ijk} is the measurement error.

The parameters to be estimated are the grand mean, the between-laboratory standard deviation s_L, the between-bottle standard deviation s_bb, and the repeatability standard deviation s_r. They are related as follows to the error terms:

$$s_\text{L} = \sqrt{\text{Var}(A_i)}$$

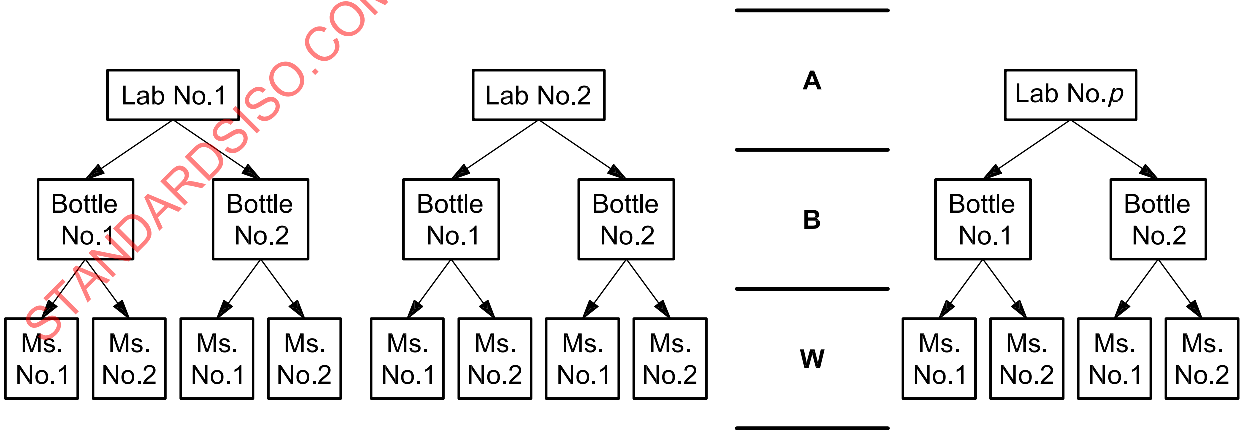
$$s_\text{bb} = \sqrt{\text{Var}(B_{ij})} \qquad \text{(A.5)}$$

$$s_r = \sqrt{\text{Var}(\varepsilon_{ijk})}$$

For the between-bottle homogeneity (s_bb), the same considerations with respect to the inability to detect batch inhomogeneity as for the homogeneity study itself (see 7.9) apply.

All these parameters can be estimated simultaneously by the analysis of variance (ANOVA) method [20] if there are sufficient results of equal replication (the same number of replicate determinations from each unit and the same number of units per method/laboratory) after any technically or statistically invalid results have been excluded. If this ANOVA requirement cannot be met because of the number of invalid and/or missing results, the significance of the between-bottle variance should be determined by other means (see Clause 7).

Theoretical details and additional methods for balanced and unbalanced ANOVA are given in standard textbooks [44], [45]. A discussion of ANOVA in the context of the certification of reference materials is given in the literature [18], [20], [22], [23].



Key

A between-laboratory variation
B between-bottle variation
W measurement repeatability

Figure A.1 — Outline of a collaborative study, combined with batch homogeneity study
[characterization of an RM (2-way layout)]

The formulae for computing the above-mentioned estimates read as follows (see References [20] and [21]). The grand mean is computed using

$$\overset{=}{x} = \frac{1}{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}} \sum_{i=1}^{p}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}} x_{ijk} \qquad \text{(A.6)}$$

where p denotes the number of laboratories, b_i the number of bottles used by method/laboratory i, and n_{ij} is the number of replicate measurements on bottle ij. The variances are computed as follows

$$\text{Var}\left(\varepsilon_{ijk}\right) = MS_{\text{within}} \qquad \text{(A.7)}$$

$$\text{Var}\left(B_{ij}\right) = \frac{MS_{B\subset A} - MS_{\text{within}}}{n_0} \qquad \text{(A.8)}$$

$$\text{Var}\left(A_i\right) = \frac{MS_{\text{among}} - n_0^{'}\text{Var}\left(B_{ij}\right) - \text{Var}\left(\varepsilon_{ijk}\right)}{(nb)_0} \qquad \text{(A.9)}$$

where

$$MS_{\text{among}} = \frac{\sum_{i=1}^{p} n_i \left(\bar{x}_A - \overset{=}{x}\right)^2}{p-1} \qquad \text{(A.10)}$$

$$MS_{B\subset A} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}\left(\bar{x}_B - \bar{x}_A\right)^2}{\sum_{i=1}^{p} b_i - p} \qquad \text{(A.11)}$$

$$MS_{\text{within}} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}}\left(x_{ijk} - \bar{x}_B\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij} - \sum_{i=1}^{p} b_i} \qquad \text{(A.12)}$$

and

$$n'_0 = \frac{\sum_{i=1}^{p}\left(\frac{\sum_{j=1}^{b_i} n_{ij}^2}{\sum_{j=1}^{b_i} n_{ij}}\right) - \frac{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}}}{p-1} \qquad \text{(A.13)}$$

$$n_0 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij} - \sum_{i=1}^{p}\left(\frac{\sum_{j=1}^{b_i} n_{ij}^2}{\sum_{j=1}^{b_i} n_{ij}}\right)}{\sum_{i=1}^{p} b_i - p} \qquad \text{(A.14)}$$

$$(nb)_0 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij} - \frac{\sum_{i=1}^{p}\left(\sum_{j=1}^{b_i} n_{ij}\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}}}{p-1} \qquad \text{(A.15)}$$

The mean squares (MS) can also be obtained using a common spreadsheet program or statistical software package. The expressions given account for missing and/or removed (invalid) data. For complete data sets, the simpler formulae of ISO 5725-3 [3] may be used.

A.3 Nested random effects in data analysis: One-way ANOVA

This model is used when the between-bottle homogeneity is verified by other means (see Clause 7). The experimental scheme is illustrated Figure A.2. The results can then be simplified to

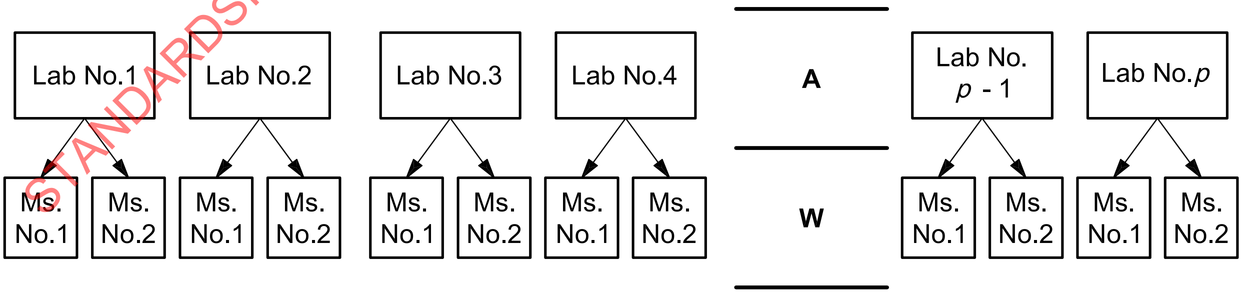
$$x_{ij} = \mu + A_i + \varepsilon_{ij} \qquad \text{(A.16)}$$

where

x_{ij} is the jth result of method/laboratory i;

A_i is the error due to method/laboratory i;

ε_{ij} is the measurement error.



Key

A between-laboratory variation
W measurement repeatability

Figure A.2 — One-way analysis of variance: lay-out of a measurement campaign collaborative study
[characterization of an RM (1-way layout)]