GUIDE 35

Fourth edition
2017-08

Reference materials — Guidance for characterization and assessment of homogeneity and stability

Matériaux de référence — Lignes directrices pour la caractérisation et l'évaluation de l'homogénéité et de la stabilité

Reference number
ISO GUIDE 35:2017(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html

This document was prepared by Technical Committee ISO/REMCO, the Reference Materials Committee of ISO.

This fourth edition cancels and replaces the third edition (ISO Guide 35:2006).

Introduction

The production of reference materials (RMs) is a key activity for the improvement and maintenance of a worldwide coherent measurement system. As detailed in ISO Guide 33[1], RMs with different characteristics are used in measurements, such as calibration, quality control, proficiency testing and method validation, as well as for the assignment of values to other materials. Certified reference materials (CRMs) are also used to confirm or establish metrological traceability to conventional scales, such as the octane number, hardness scales and pH.

To be comparable across borders and over time, measurements need to 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 substances and materials. 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 the International System of Units (SI) or other internationally agreed references during production. This document explains how approaches can be developed that will lead to well established property values, which are made traceable to appropriate stated references.

For reference material producers (RMPs), there is an International Standard and three ISO Guides that support the production and certification of RMs to ensure that the quality of the RMs meets the requirements of the end users.

- ISO 17034 outlines the general requirements to be met by an RMP to demonstrate competence.
- ISO Guide 35 provides more specific guidance on technical issues and explains the concepts for processes such as the assessment of homogeneity, stability and characterization for the certification of RMs.
- ISO Guide 31[2] describes the contents of certificates for CRMs, and of accompanying documents for other RMs, respectively.
- ISO Guide 30[68] contains the terms and definitions related to reference materials.

Alongside developments in RM production approaches, the range of classes of RMs is growing with advances in technology, increasing the need for more widely applicable technical guidance in RM production. In addition, increasing use of ISO/IEC 17025[52] and ISO 15189[71] by laboratories has led to greater demand for clear statements of metrological traceability.

This document provides detailed guidance on a larger range of homogeneity study designs, and describes a wider range of stability management strategies than ISO Guide 35:2006. It also contains specific provisions concerning the establishment of metrological traceability in RM production.

Reference materials — Guidance for characterization and assessment of homogeneity and stability

1 Scope

This document explains concepts and provides approaches to the following aspects of the production of reference materials:

— the assessment of homogeneity;

— the assessment of stability and the management of the risks associated with possible stability issues related to the properties of interest;

— the characterization and value assignment of properties of a reference material;

— the evaluation of uncertainty for certified values;

— the establishment of the metrological traceability of certified property values.

The guidance given supports the implementation of ISO 17034. Other approaches may also be used as long as the requirements of ISO 17034 are fulfilled.

Brief guidance on the need for commutability assessment (6.11) is given in this document, but no technical details are provided. A brief introduction for the characterization of qualitative properties (9.6 to 9.10) is provided together with brief guidance on sampling such materials for homogeneity tests (Clause 7). However, statistical methods for the assessment of the homogeneity and stability of reference materials for qualitative properties are not covered. This document is also not applicable to multivariate quantities, such as spectral data.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 3534-2, *Statistics — Vocabulary and symbols — Part 2: Applied statistics*

ISO 3534-3, *Statistics — Vocabulary and symbols — Part 3: Design of experiments*

ISO Guide 30, *Reference materials — Selected terms and definitions*

ISO/IEC Guide 99, *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*

NOTE The *International vocabulary of metrology* will hereafter be referred to as the "VIM".

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO Guide 30, ISO/IEC Guide 99, ISO 3534-2, ISO 3534-3 and the following apply. The definitions in ISO Guide 30 take precedence where more than one definition for the same term exists.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

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 to entry: RM is a generic term.

Note 2 to entry: Properties can be quantitative or qualitative, e.g. identity of substances or species.

Note 3 to entry: Uses may include the calibration of a measurement system, assessment of a measurement procedure, assigning values to other materials, and quality control.

Note 4 to entry: ISO/IEC Guide 99:2007[3] has an analogous definition (5.13), but restricts the term "measurement" to apply to quantitative values. However, ISO/IEC Guide 99:2007, 5.13, Note 3 (VIM), specifically includes qualitative properties, called "nominal properties".

[SOURCE: ISO Guide 30:2015, 2.1.1]

3.2

certified reference material

CRM

reference material (RM) characterised by a metrologically valid procedure for one or more specified properties, accompanied by an RM certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability

Note 1 to entry: The concept of value includes a nominal property or a qualitative attribute such as identity or sequence. Uncertainties for such attributes may be expressed as probabilities or levels of confidence.

Note 2 to entry: Metrologically valid procedures for the production and certification of RMs are given in, among others, ISO 17034 and ISO Guide 35.

Note 3 to entry: ISO Guide 31[2] gives guidance on the contents of RM certificates.

Note 4 to entry: ISO/IEC Guide 99:2007[3] has an analogous definition (5.14).

[SOURCE: ISO Guide 30:2015, 2.1.2]

3.3

measurement model

mathematical relation among all quantities known to be involved in a measurement

[SOURCE: ISO/IEC Guide 99:2007, 2.48[3]]

3.4

property value

<of a reference material (RM)> value corresponding to a quantity representing a physical, chemical or biological property of an RM

[SOURCE: ISO Guide 30:2015, 2.2.1]

3.5

certified value

value, assigned to a property of a reference material (RM), that is accompanied by an uncertainty statement and a statement of metrological traceability, identified as such in the RM certificate

[SOURCE: ISO Guide 30:2015, 2.2.3]

3.6
indicative value
information value
informative value
value of a quantity or property of a reference material, which is provided for information only

Note 1 to entry: An indicative value cannot be used as a reference in a metrological traceability chain.

[SOURCE: ISO Guide 30:2015, 2.2.4]

3.7
calibrant
reference material used for calibration of equipment or a measurement procedure

[SOURCE: ISO Guide 30:2015, 2.1.21]

3.8
quality control material
reference material used for quality control of a measurement

[SOURCE: ISO Guide 30:2015, 2.1.22]

3.9
isochronous stability study
experimental study of reference material stability in which units exposed to different storage conditions and times are measured in a short period of time

3.10
production
<of a reference material (RM)> all necessary activities and tasks leading to the release and maintenance of an RM (certified or non-certified)

Note 1 to entry: Activities include, for example, planning, control, material handling and storage, material processing, assessment of homogeneity and stability, characterization, assignment of property values and their uncertainties, authorization and issue of RM certificates or other statements.

[SOURCE: ISO Guide 30:2015, 2.3.7]

4 Symbols

a	number of reference material units in a homogeneity study
d	measurement bias
k	coverage factor or (as subscript) index
L_d	a limit of detection (minimum detectable value of the net state variable) calculated using the methods of ISO 11843-1[48]
N_min	minimum number of RM units for a homogeneity study for batch sizes over 100 units
N_prod	number of RM units produced in a single batch
n_r	number of runs in a blocked or nested homogeneity study design
p	number of laboratory means in an interlaboratory certification exercise
s_bb	between-unit component of variance from a homogeneity study, expressed as a standard deviation

s_r	repeatability standard deviation
s_R	reproducibility standard deviation
t_lts	duration of a long term stability study
U_CRM	expanded uncertainty associated with a property value of the CRM
u_bb	standard uncertainty associated with between-unit variability
u_char	standard uncertainty associated with a value assigned in a characterization study
u_CRM	standard uncertainty associated with property value of the CRM
u_trg	target measurement uncertainty, expressed as standard uncertainty, for the value of a property to be certified
u_hom	standard uncertainty associated with heterogeneity
u_lts	standard uncertainty associated with long term stability
u_mon	standard uncertainty associated with a value obtained by measuring an RM at a monitoring point
u_trn	standard uncertainty associated with the transport stability of the material
u_wb	standard uncertainty associated with within-unit heterogeneity
x_CRM	property value of a CRM
$\hat{x}$	estimated value obtained from a robust statistical estimator
x_mon	value obtained by measuring an RM property value at a monitoring point
x	amount-of-substance fraction
y_char	value assigned to a reference material in a characterization study

Additional symbols used in particular subclauses are defined on first use in the text.

5 Conventions

In this document, the following conventions are used.

a) A measurand is specified in such a way that there exists a unique 'true value'.

b) All probability assessments described in this document assume normality unless otherwise stated.

c) Throughout this document, the law of propagation of uncertainty is used for the combination of measurement uncertainty contributions. Other methods of evaluating measurement uncertainty may also be applied, and in some cases it is necessary to do so. Further guidance on these matters is given in ISO/IEC Guide 98-3, "*Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*" and its supplements (see References [5] and [6]).

NOTE 1 Variation between units associated with heterogeneity and changes due to instability might not be normally distributed and can result in asymmetric distributions.

NOTE 2 The "*Guide to the expression of uncertainty in measurement*" will hereafter be referred to as the "GUM".

6 An overview of reference material production

6.1 General

The production and distribution of an RM require careful planning prior to undertaking any actual activity in the project. The following subclauses provide a brief overview of the steps involved in the production of a reference material followed by a description of the main issues involved in planning each step. Detailed guidance on homogeneity assessment, stability assessment and characterization is given in Clauses 7, 8 and 9, respectively.

6.2 Summary of project design

The production of a reference material involves the following steps:

a) definition of the RM, i.e. the matrix, the properties to be characterized and their desired levels, the intended use of the material, and for CRMs, the target uncertainty[72];

b) design of a procedure for the sourcing of the material;

c) design of a reference material manufacturing and/or preparation procedure;

d) selection of measurement procedures appropriate for characterization, homogeneity and stability studies;

e) consideration of metrological traceability for each measured property, particularly for CRMs, for which a statement of metrological traceability is required;

f) assessment of homogeneity;

g) assessment of stability;

h) assessment of commutability (if required);

i) characterization of the reference material;

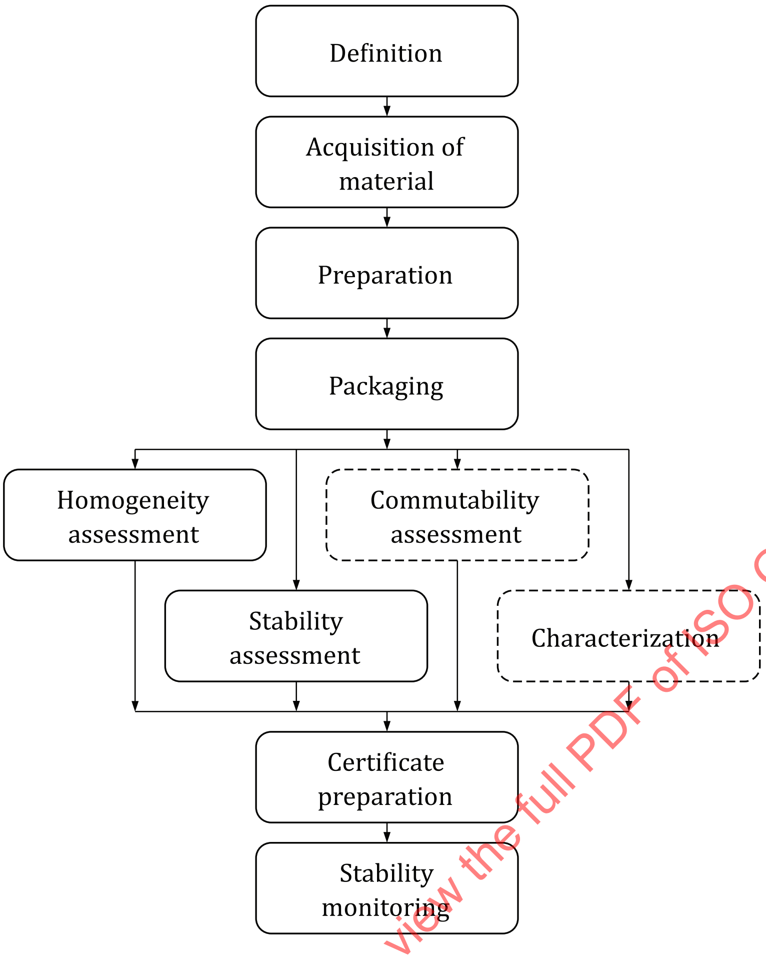
j) combination of the results from homogeneity studies, stability studies, and, for CRMs, evaluation of the measurement uncertainties of certified values;

k) preparation of a certificate or product information sheet and, if appropriate, a report on the production and/or certification;

l) specification of storage and transportation conditions;

m) post-production monitoring of stability.

The main stages are shown schematically in Figure 1.



NOTE 1 The figure provides a schematic outline of the main steps in producing and maintaining a reference material. Boxes with dashed outlines are not always necessary.

NOTE 2 'Packaging' in this diagram includes subdivision into individual units in suitable containers for distribution.

NOTE 3 In this diagram, 'Certificate preparation' includes all types of documentation that could be provided with a reference material, including a certificate, product information sheet, certification report, etc.

Figure 1 — Schematic outline of a reference material project

6.3 Acquisition of starting material

The first task in an RM production project is the acquisition of a sufficient amount of starting material(s) with the desired properties. The production of materials with particular properties is considered briefly in 9.3.4. The amount of material needed is determined by the following:

- the number of units of the RM needed for distribution over the expected life of the RM;
- the number of units needed for the homogeneity study;
- the number of units needed for the stability study;
- the number of units needed for the characterization of the candidate RM;
- the number of units required for monitoring stability over the expected lifetime of the material;
- the planned size of each RM unit, which has to be sufficient for at least one measurement;
- the need for one or more feasibility studies;

- optionally, additional units to cover contingencies such as, for example, follow-up studies to respond to customer queries, future recertification required by a significant change in the storage conditions, or extension of the number of certified properties.

The number of units of a candidate RM that are needed for distribution is often, at least in part, a commercial issue and should be carefully considered before commissioning the collection and processing of the material. In addition, the expected long-term stability of the material in storage can influence the amount of material that can usefully be produced. It may be prudent to limit the number of units produced for less stable materials to avoid wastage due to unavoidable degradation over time.

6.4 Feasibility studies

Feasibility studies are short studies intended to address concerns about the feasibility of producing and characterizing a sufficiently homogeneous and stable RM. For example, questions such as the best way of preparing the RM or ensuring sufficient stability of the material can be answered by small-scale feasibility studies early in the project[7].

Where characterization is expected to be performed through the use of an interlaboratory study, a feasibility study can identify possible sources of error and enable participants involved in the characterization to optimize their equipment and procedures.

NOTE In a feasibility study intended to test or improve the capabilities of participants in an interlaboratory characterization exercise (see Clause 9), use of a material different from the candidate RM can avoid undue bias in participant results arising from prior knowledge of the candidate RM.

6.5 Reference material processing

Processing can involve a range of processes, including, for example:

- synthesis, manufacture or formulation of a synthetic reference material;
- drying, lyophilisation, milling, and/or filtration for natural materials;
- addition of stabilizing agents;
- homogenization prior to packaging.

The particular procedures used depend on the particular material and usually require expert guidance.

6.6 Homogeneity assessment

Homogeneity is an important requirement for all RMs and includes both within- and between-unit homogeneity. Between-unit homogeneity is important to ensure that each RM unit carries the same value for each property; within-unit homogeneity is important where subsamples can be taken for measurement by users of the material. Clause 7 gives detailed guidance on homogeneity assessment.

6.7 Stability assessment

RMs should be sufficiently stable for their intended use, so that the end user can rely on the assigned value at any point within the period of validity of the certificate. Typically, it is important to consider stability under long-term storage conditions, under transport conditions and, where applicable, the storage conditions at the RM user's laboratory. This can include consideration of stability after opening, if re-use is permitted. Clause 8 provides detailed guidance on stability assessment.

6.8 Choice of measurement procedures

In a reference material production project, each step that requires measurements may use different measurement procedures because, for example, characterization generally requires minimally biased measurement procedures with low uncertainty; homogeneity studies primarily require the best available repeatability; and classical stability studies typically require measurement procedures that show good precision over time within the same laboratory. The choice of measurement procedures for homogeneity studies, stability studies and characterization is considered in Clauses 7, 8 and 9, respectively.

6.9 Metrological traceability

Metrological traceability is key to ensuring the comparability of measurement results over time and between locations, including those used to characterize reference materials. By definition, CRMs are accompanied by a statement of metrological traceability for each certified property value. The proper choice of the stated references to which metrological traceability of the property values is established is essential for CRMs, because CRMs are primarily used to make measurement results traceable. Establishment of metrological traceability is considered in detail in 9.2.

6.10 Characterization and uncertainty evaluation

Characterization refers to the determination of the property values of the relevant properties of an RM, as part of the production process. Characterization of an RM is described in Clause 9. For CRMs, certified values are accompanied by a statement of measurement uncertainty; the evaluation of uncertainty is considered in Clause 10.

6.11 Commutability assessment

The commutability of an RM relates to the ability of the RM, characterized by one measurement procedure (usually a reference procedure) to act as a calibrator or quality control (QC) material for a second measurement or testing procedure applied to routine test materials. This is particularly important where different measurement procedures can respond very differently to different types of test materials. Commutability assessment is not required for all RMs but is required for some important classes of RM.

NOTE Current ISO/REMCO information on commutability assessment[9] states that:

“A reference material producer should conduct an assessment of commutability where

a) the intended use requires commutability of calibration or quality control materials,

and

b) the reference material producer warrants that the material is fit for the intended use.

NOTE 1 Demonstration of commutability is usually required when the intended use includes calibration or quality control in biological measurement, and is not usually required when the intended use does not include biological measurement and the procedure is known to be adequately specific for the measurand in the matrix of the reference material and the intended routine samples.

NOTE 2 It is not usually necessary to establish commutability when the reference material and its origin are obtained from sources and handled the same as samples that would be tested for customers, for example, matrix reference materials.

6.12 Transport issues

Nearly all RMs have to be transported to the location of use. The means and conditions of transport of an RM after production are relevant to the need for stability studies (see Clause 8) and it is therefore useful to consider transport conditions at an early stage in the project.

NOTE 1 National and/or international transport regulations may limit the options for transport, prohibit the transport of some materials, or require specific packaging or precautions for safety or other reasons.

NOTE 2 The time taken for official procedures such as customs or other border control clearance can increase shipment times for some destinations.

6.13 Value assignment

Value assignment is the process of combining the results from the homogeneity and stability assessment with the results from the characterization studies to determine the assigned values and their uncertainties. These values are subsequently issued on a certificate or product information sheet.

6.14 Stability monitoring

Most reference materials are stored for extended periods at the RM producer's premises or by distributors. Since stability assessment cannot usually anticipate all changes that may occur, it is usually necessary, as a part of managing the risks associated with possible instability, to monitor the property values of materials held for extended periods. Because the requirements for monitoring depend in part on knowledge obtained during stability assessment, 8.10 includes guidance on stability monitoring.

6.15 Reference materials produced in repeated batches

The need for experimental study of some characteristics (particularly homogeneity, stability and commutability) can be reduced where the material is produced in a repeat production run following an established procedure. Reliance on prior experience is reasonable so long as:

a) the process for producing batches of the RM has not changed in any way that might adversely affect the end use;

b) the materials used in production of the RM have not changed in any way that might adversely affect the end use;

c) materials previously produced by the same process have shown no failures attributable to the production process, either during routine monitoring or by users; and

d) the requirements for the material are reviewed regularly, taking account of the intended use of the material at the time of the review, to ensure that the production process remains fit for purpose.

Consistent performance of the production process should be checked, for example by comparing the property values of samples from successive batches under repeatability conditions.

7 Assessment of homogeneity

7.1 Preamble

Most RMs are prepared as batches of 'units' (e.g. bottles, vials or test pieces). It is important that all distributed units are the same within the stated uncertainty for each property value and, unless sold as single-use units, that the material within each unit is uniform. ISO 17034 accordingly requires the assessment of the homogeneity of a reference material (RM).

Homogeneity can refer either to variation of a property value between separate units of the material, or to variation within each unit. It is always necessary to assess the between-unit variation. Where the intended use permits the use of part of a unit – for example, a small portion of a solid or liquid material, or a small region of the surface – it is also usually necessary either to assess the within-unit variability of the material (within-unit heterogeneity) or to provide instructions for use that control the impact of within-unit heterogeneity. These instructions can include, for example, remixing of the sample and, for granular materials, a minimum sample size, because the within-unit heterogeneity is directly reflected in the minimum size of subsample that is representative for the whole unit.

The assessment of homogeneity may include the use of prior evidence (including prior experimental evidence) of the homogeneity of the material, performing an experimental homogeneity study on the candidate reference material, or both. In most cases, an experimental study is necessary. Exceptions include, for example, batches of a highly homogeneous material, such as a solution for which previous experimental studies have demonstrated that packaging and storage do not affect the homogeneity; or the production of materials for which each reference material unit has a separate assigned value. 7.2 gives further details about the circumstances and types of material requiring experimental study.

The results of an experimental homogeneity study are usually also used for the calculation of one of the uncertainty components in the certification model (see Clause 10). The magnitude of this uncertainty component can vary widely compared with other components of the uncertainty, depending on the nature of the RM and of the certified property.

To undertake a homogeneity study, a subset of units, typically 10 to 30, is chosen from the batch using a suitable sampling scheme, property values are measured for each unit using a suitable measurement procedure and the results are assessed using appropriate statistical methods to obtain information on, for example, between-unit variability and within-unit variability of the material.

To obtain reliable results, it is important to

- choose the properties to be studied,
- select a representative subset of units,
- choose a suitable measurement procedure with sufficient repeatability and selectivity,
- make the measurements under suitable conditions following an appropriate experimental design, and
- conduct the statistical analysis using valid statistical methods.

7.3 to 7.7 provide guidance on each of these steps. Examples of calculations are provided in Annex C.

Historically, homogeneity studies have tested for statistically significant between-unit differences compared with measurement precision in order to decide whether a material is homogeneous or not. This approach is not taken in this document; rather, emphasis is placed on deciding whether the between-unit standard deviation is sufficiently small for the intended end use. Statistical tests of significance may, however, be of use in RM production, for example in order to decide whether further processing is required to reduce heterogeneity to insignificance compared with routine measurement precision. If statistical tests are used, however, the homogeneity experiment should be capable of detecting important heterogeneity, in turn requiring a sufficient combination of precision of the measurement procedure, number of RM units and number of replicates per unit. Statistical power calculations (7.4.2) can assist in ensuring a sufficiently effective test.

7.2 Need for an experimental homogeneity study

Materials of natural origin or with complex matrices, such as foodstuffs, soils, ores and alloys, are typically heterogeneous in composition. Although the magnitude of between-unit differences can sometimes be small or even negligible after homogenization, in other cases, between-unit differences can remain larger than the uncertainty arising from characterization. RMs prepared from such heterogeneous materials should therefore be subjected to an experimental homogeneity study.

RMs prepared as pure compounds or solutions of pure compounds (if certified for purity; not for impurities) are expected to have a high degree of homogeneity. These materials can, however, also show some heterogeneity, for example, due to a density gradient, localized contamination, evaporation of solvent during processing or filling, variations in residual solvent content, or metals containing variable amounts of occluded gases. Furthermore, certified values for such materials are often expected to have very small uncertainties, making even a small amount of heterogeneity potentially important. Even in cases where the material is expected to be sufficiently homogeneous for most intended uses, homogeneity should be verified. Verification may include a complete homogeneity study or other check (for example, a check on melting point consistency between units of a pure organic material).

An experimental study of the homogeneity of a material is not essential in the following cases:

- when the material is a repeat production run of a previous material that was produced following the same procedure and that has been shown by experiment to be sufficiently homogeneous; or

— where the production process has been validated and thereby shown to consistently produce sufficiently homogeneous batches of material.

Examples of materials that are sometimes produced in this way include ethanol calibration solutions or elemental calibration solutions prepared by mass and thoroughly mixed to ensure that the mixture will be sufficiently homogeneous for the intended use.

Where assurance of homogeneity relies on a validated production process, quality control procedures should be used to confirm consistent operation of the production process. Such procedures may include, for example, operation of a range control chart or standard deviation control chart for monitoring the range or standard deviation of a small number of units measured, or criteria for the range of values found in each characterization.

NOTE 1 6.15 provides additional guidance regarding reliance on experience gained from previous production batches.

NOTE 2 ISO 7870-2[73] gives guidance on the use of range control charts and standard deviation control charts.

7.3 Properties to be studied

Where homogeneity is to be determined experimentally it is usually necessary to determine the homogeneity for every property of interest, that is, every property for which the material is claimed to be sufficiently homogeneous for the intended use.

The homogeneity of all properties of interest may be assessed by examining a smaller number of selected properties when

— there is sufficiently high statistical correlation between particular property values in the type of material of interest to allow useful prediction of one property value from one or more others, and

— it can be shown that particular groups of properties are sufficiently closely associated (for example, because of their presence in a particular component of a mixture) that measurement of one property in such a group furnishes evidence of homogeneity for other properties in the same group.

It is essential that any subset of properties taken as representing homogeneity for a larger set of properties 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 would support the assumption that the RM also has a similar degree of homogeneity for the non-selected elements.

In cases where homogeneity is assessed experimentally by using a subset of properties of interest, additional evidence should be gained about the homogeneity of properties that are not studied experimentally. The evidence should be sufficient to show that the uncertainty associated with heterogeneity is not underestimated for those properties that are not experimentally studied.

NOTE Such evidence can be gained, for example, from literature relevant to the material in question, through the stability study or the characterization of the material.

7.4 Statistically valid sampling schemes

7.4.1 Minimum number of units for a homogeneity study

7.4.1.1 Homogeneity study for quantitative properties

Homogeneity studies for quantitative properties are typically intended to provide information on the variance due to heterogeneity and on any (possibly nonlinear) trends arising from processing. To achieve these objectives, the number of items should be sufficient to give a reasonable estimate of the between-unit variance, and sufficient items taken to give a clear view of any trends present.

Based on current practice, an acceptable estimate of the between-unit variance for the purposes of uncertainty evaluation can be obtained with nine or more degrees of freedom. For a simple homogeneity study design or a randomized block design (see below), this corresponds to the selection of a minimum of 10 units. Where a nested design is to be used for measurements, in which a subset of units is measured in each of several runs, additional units should be included in the study in order to maintain the required degrees of freedom. 7.6.4 gives further details about the additional number of units required for a nested design.

Trends arising from processing often appear as an initial trend followed by stable output, as a trend developing late in the process, or as a combination of the two. Where these features occur, it is often possible to provide a homogeneous material by discarding the affected units from the beginning or end of the run. However, examining only 10 units might not provide sufficient information about the onset of trends near the ends of a lengthy processing run. Current best practice therefore increases the number of units examined as the total number of units produced, N_{prod}, increases. Typical recommendations are between $\sqrt[3]{N_{\text{prod}}}$ and $3\times\sqrt[3]{N_{\text{prod}}}$ [74]. Taken together with the degrees of freedom requirement above, this leads to a recommended minimum number of units N_{min} for a homogeneity study of materials characterized for a quantitative property given by

$$N_{\text{min}} = \max\left(10, \sqrt[3]{N_{\text{prod}}}\right) \tag{1}$$

where max(.., ..) indicates the maximum of the terms within the parentheses.

NOTE 1 It is not normally useful to examine more than 30 units of a reference material characterized for a quantitative property.

NOTE 2 7.4.1.3 gives further guidance on the minimum number of units for production batches of 100 or fewer units

EXAMPLE An RM producer prepares 3 000 units of a candidate RM and intends to undertake a homogeneity study in a single run. Then N_{prod} = 3 000 and $N_{\text{min}} = \max\left(10, \sqrt[3]{3\,000}\right)$, which is 14,4. This study accordingly requires 15 units of material for the homogeneity study.

7.4.1.2 Homogeneity study for qualitative (“nominal”) properties

For reference materials certified for a qualitative, or “nominal”, property, the number of units chosen for the homogeneity study should be set based on sampling guidance for inspection by attributes as described in the ISO 2859 series[8] or similar guidance.

Sampling plans for inspection by attributes lead to very high inspection numbers if low proportions of defective units are to be detected by sampling alone. For example, to detect a 1 % defective rate with high confidence based on one or more observed defectives, about 300 randomly chosen units have to be inspected (this number is based on a statistical power calculation for 95 % test power for 1 % defectives, assuming batch size much greater than 300 units). This is often unrealistic for typical reference material batch sizes. For qualitative materials, therefore, an experimental homogeneity study is likely to be limited to a check for unexpected gross heterogeneity (for example, greater than 10 % defective), with other information on origin of material, processing, etc. used to support any statement of homogeneity.

If a certified qualitative property is individually verified for every unit of such a material, it is not necessary to perform a further test of the homogeneity of that property.

7.4.1.3 Small production batches

Some reference materials are produced in small batches of 50 or fewer units, for example, secondary gas calibration standards. For such small production batches, the minimum number of units specified in 7.4.1.1 usually represents a very large fraction of the available units. Where the batch size is below 100 units, homogeneity should be assessed on the larger of three units or 10 % of the batch size, randomly selected from the batch. Replication should be as high as practically feasible to provide the best available test power for the number of units used. Power analysis (7.4.2) may be used to assist in considering desirable replication levels. For example, with three units, four observations per unit gives approximately the same power of detecting normally distributed between-unit deviations as observing 10 units in duplicate, though with much lower probability of detecting occasional defective units in an otherwise homogeneous batch.

7.4.2 Use of statistical power analysis

Statistical power analysis may be used to assist in choosing a suitable number of units and replicates for the homogeneity study. Power analysis aims to control the probability of failing to detect a particular level of heterogeneity given a proposed statistical test for significant heterogeneity. Power analysis is a specialized topic but is increasingly available in software, some available without charge. The most common example of power analysis in this context is the calculation of the numbers of test items and replicate measurements on the assumption that one-way analysis of variance (see 7.7.3) is to be used to test for a significant between-unit effect.

Although statistical power analysis can be useful in comparing different proposed strategies, considerable care should be taken in its use. In particular, the choice of replicate numbers and unit numbers is very strongly dependent on the assumed distribution of any hypothesized between-sample difference. For example, assuming a normal distribution for the (true) means of different units, a common default in power calculation software, leads to a high proposed number of replicates and a small number of units. This is a poor choice if the most likely pattern of heterogeneity is a small proportion of discrepant units among a largely homogeneous population. In the absence of good information on the likely distribution of different (true) unit means, therefore, power analysis is most useful as an aid in choosing replicate numbers after the proposed number of units has been decided.

NOTE 1 When using power analysis, a typical choice of test significance and intended power is 95 % and 80 %, respectively, though note that this provides substantially greater power than the minimum number of units and replicates considered at 7.4.1.1.

NOTE 2 Details of some simple power calculations appropriate for reference material studies are given in Reference [57].

7.4.3 Sampling strategy for a homogeneity study

The sampling scheme used to pick the units (items, bottles) for a homogeneity study is typically one of simple random sampling, stratified random sampling or systematic sampling. The general procedures for each, as applied to reference material sampling for homogeneity studies, are as follows:

— Simple random sampling takes a random sample from the batch.

— Random stratified sampling typically divides the batch into a number of segments of equal size (usually by production or packaging sequence or, sometimes, location) and takes an equal number of units (often one) at random from each such segment.

— Systematic sampling sets an interval n_{syst} between sampled units and a random start point n_1 between 1 and n_{syst}, and takes units at n_1, n_1+n_{syst}, n_1+2n_{syst} ..., etc. n_1 and n_{syst} are chosen to give complete coverage of the production batch and the desired number of units.

Random samples should be defined using random number generation software or random number tables. Manual arbitrary choice is not equivalent to random selection.

Simple random sampling is most appropriate when there is no known ordering, or for small batches in which a large proportion of the units is used for homogeneity study. Stratified random sampling is recommended for most situations, since this guarantees that the units picked for the homogeneity study are approximately evenly distributed over the whole batch and avoids coincidence between sampling location and any cyclic effects in processing. Systematic sampling may be applied when there is little risk of overlooking repetitive effects or trends in the batch.

The sampling scheme should take into consideration potential weaknesses in the method of preparing and storing samples (including, for example, processing and fill trends and possible settling on storage prior to subdivision), thus allowing a critical examination of the prepared batch.

Units should normally be numbered before sampling, or numbered in processing order after sampling, to permit subsequent trend analysis.

Alternative sampling schemes may be used where it can be shown that the resulting variance estimates are not biased by the sampling scheme chosen.

7.5 Choice and conduct of the measurement procedure for a homogeneity study

7.5.1 Choice of measurement procedure

The measurement procedure chosen to make measurements for homogeneity studies should be chosen primarily for good precision during the expected duration of each measurement run (that is, a good repeatability standard deviation, s_r) and, if units are to be randomized among several runs, good between-run precision. The standard deviation for measurements should be small compared with the expected uncertainty for the value for each property measured; ideally, the repeatability standard deviation for the homogeneity study procedure should be less than one third of the desired standard uncertainty. That is, if the target measurement uncertainty for the property value is u_{trg} (expressed as standard uncertainty), the repeatability standard deviation for the procedure should ideally comply with

$$\frac{s_r}{\sqrt{n_{\text{al}}}} \le \frac{u_{\text{trg}}}{3} \qquad (2)$$

where n_{al} is the number of observations on each of n aliquots taken from each unit for measurement.

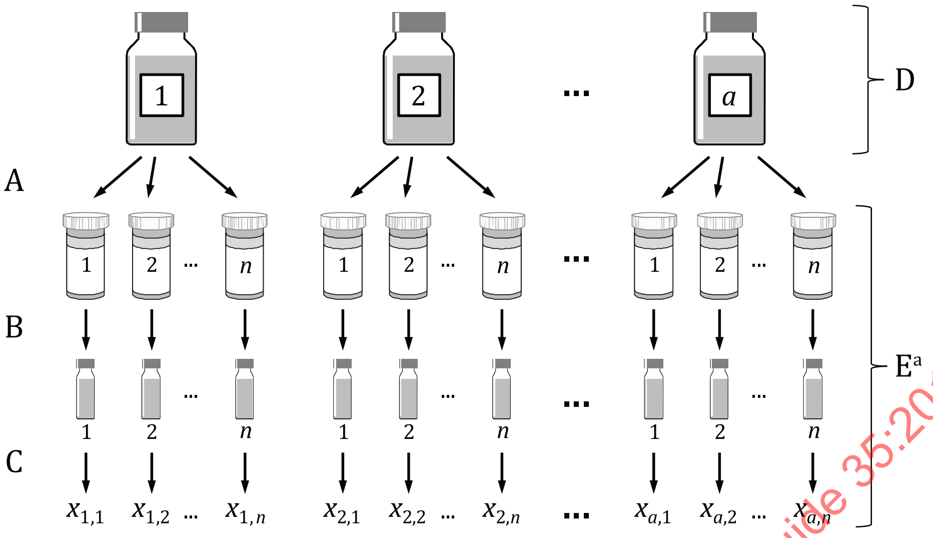
Where this precision requirement cannot be met in practice, for example where RM units do not allow high replication, every effort should be made to ensure that $s_r/\sqrt{n_{\text{al}}}$ does not exceed the target standard uncertainty u_{trg}, and consideration should be given to increasing the number of units.

NOTE 1 Increasing the number of units is not as effective for overcoming the effects of poor precision as is increasing the number of observations per unit or improving measurement precision, unless the focus is on finding infrequent, highly discrepant units. Statistical power calculations (see 7.4.2) can be of use in choosing a suitable number of units and replicates to give an experimental design that provides similar assurance to that obtained when the precision requirements in Formula (2) are met.

NOTE 2 An additional consideration in the choice of a measurement procedure is the amount and complexity of sample preparation needed to get each specimen into measurable form. Procedures requiring little or no sample preparation (so-called non-destructive methods) can show better precision than procedures requiring extensive, multi-stage, sample preparation.

7.5.2 Conduct of measurements for homogeneity studies

It is important, where possible, to compare the dispersion of results for different RM units with the dispersion under repeatability conditions (7.6.1). “Repeatability conditions” involve replication of the complete measurement procedure. It is therefore important that, wherever possible, the complete sample preparation (“transformation”) and final measurement step be applied to each subsample or aliquot taken from each unit (see Figure 2). If this is not done, any variation in sample preparation prior to replicate observations will inflate the estimated between-unit term (Figure 3), resulting in an over-estimate of the uncertainty associated with heterogeneity.



Key

A subsampling

B preparation

C measurement

D contributes to the observed between-unit variation

E operations contributing to observed within-unit variation

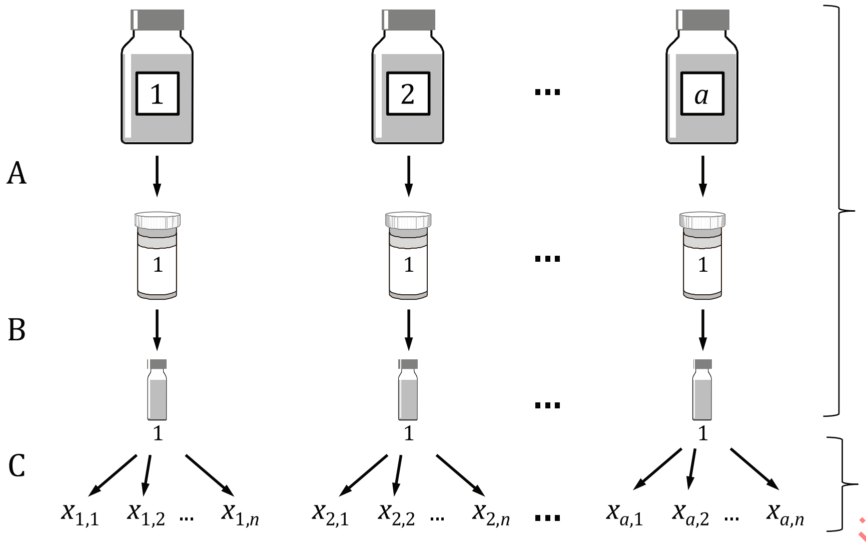
[a] $x_{i,j}$ denotes the *j*th aliquot for unit *i*

Figure 2 — Layout of a between-unit homogeneity study

In Figure 2, an ideal case is shown in which subsampling of reference material units is possible and multiple (*n*) test portions have been taken from each sample of the batch and individually prepared (transformed) for measurement. In this case, the variance within units includes the variation due to subsampling, preparation and measurement. From the perspective of obtaining an unbiased estimate of the between-unit standard deviation, this is the ideal situation because none of these effects contributes to the estimated between-unit variance.

In Figure 3, a layout is shown for the case where subsampling of the items is impossible or just not carried out; only one subsample from each of the *a* reference material units is taken and prepared for measurement, and *n* measurements are made on the prepared test portion. In this design, the effects of sample preparation and any subsampling are included in the observed between-unit variance calculated from the experiment. The calculated within-unit variance includes only the repeatability of the measurement. This procedure will result in a larger estimate of the between-unit variance, in turn over-estimating the between-unit heterogeneity.

It follows that wherever possible, homogeneity studies should be conducted so that all steps of the measurement procedure are completed on each subsample.


A
B
C
D
E
1
2
a
1
1
1
1
1
1
...
$x_{1,1}$ $x_{1,2}$ ... $x_{1,n}$
$x_{2,1}$ $x_{2,2}$ ... $x_{2,n}$
$x_{a,1}$ $x_{a,2}$ ... $x_{a,n}$

Key

A subsampling

B preparation

C measurement

D contributes to the observed between-unit variation

E operations contributing to observed within-unit variation

Figure 3 — Alternative layout of a between-unit homogeneity study

7.6 Homogeneity study designs

7.6.1 Objective of a homogeneity study

A homogeneity study uses statistical evaluation to compare the dispersion of observations on several units of the RM to the precision of the measurement procedure in order to determine the between-unit standard deviation, which can then be used to calculate the uncertainty associated with heterogeneity. This usually involves making an equal number of replicate measurements on each RM unit as illustrated in Figure 2. To obtain the best estimate of the between-unit term, the dispersion of values obtained for reference material units in a homogeneity study should be compared with the dispersion of results under repeatability conditions, wherever possible (see Reference [11]). This can be achieved by

a) conducting all homogeneity measurements for a given property in a single measurement run, or

b) using an experimental design which allows separate estimation of within-run, between-run and between-unit variances. The most useful designs for this purpose are simple randomized block designs and balanced nested designs.

The three designs considered are illustrated in Figure 4. Typical experiments using these different designs are described in more detail in 7.6.2 to 7.6.4. Data analysis is covered in 7.7.3 to 7.7.5, respectively.

a) Run 1: c a b a f e b g j d h k c h i j k f l d i l g e

b) Run 1: a e j b i d h c f l g k

Run 2: g d j f e i k l h b a c

c) Run 1: c a a h b h b c

Run 2: k i i l e e l k

Run 3: g j d d f g f j

Key

a Simple randomized design; a single run with all units observed in duplicate in random order.

b Randomized block design. Each run includes one observation on every unit, each run being separately randomized.

c Nested design. Four units are randomly allocated to each of three runs and duplicate observations are made on each, again in randomized order. Some observations from the same unit are adjacent by chance; see 7.6.2, Note 1.

All illustrations show designs for duplicate observations on 12 RM units designated from “a” to “l”.

Figure 4 — Designs for homogeneity studies

7.6.2 The basic homogeneity study design – measurement in a single run

Where the measurement system is stable for the duration of the observations required, all observations may be acquired in a single measurement run. The within-unit and between-unit standard deviations can then be evaluated using a one-way ANOVA design (see 7.7.3, B.1 and C.1). Measurements should be carried out in such a way that

i) trends in the measurement system are not misinterpreted as differences between RM units, and

ii) a trend in the measurements can, if necessary, be separated from a trend in the batch of samples.

This is best achieved by measuring the replicates of the samples used in the homogeneity study in a fully randomized order. Figure 4 a) shows an illustrative example of such a design.

RM producers may also use a systematic arrangement of replicate observations where it can be shown that

a) reasonably expected patterns of measurement error, including trends and step changes in response within the run, will not cause appreciable bias in estimates of between-unit variance, and

b) subsequent inspection and data analysis procedures are able to distinguish processing (including packaging) trends from any measurement trends during the homogeneity study.

A randomly chosen order can, by chance, lead to units ordered almost in processing order or with almost all replicates for the same RM unit close together in the measurement sequence. A proposed random ordering which produces obvious strong correlation with processing sequence or close grouping of replicates for most or all units should be rejected and a new random ordering selected.

NOTE 1 A randomized order does not guarantee that individual replicates from the same RM unit are not adjacent. Occasional instances of adjacent replicates are common and acceptable in a fully randomized design.

NOTE 2 A systematic forward-reverse ordering of observations that does not cause appreciable bias in the between-unit term in the presence of a simple linear trend in instrument response with time[12] is illustrated in the following example.

EXAMPLE Suppose 10 samples are used for a homogeneity study, with 3 replicates. A possible systematic 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

7.6.3 Randomized block design

In a randomized block design for n_0 replicates on each of a units of a reference material, the simplest randomized block design involves n_0 measurement runs and each unit is measured once in each run, in random order. Runs should be randomized individually. Figure 4 b) shows an illustrative example of such a design. Two-way analysis of variance without replication, or other methods (see 7.7.4 and 7.7.6), can then estimate the within- and between-unit standard deviations independently of the run effect.

A randomized block design is appropriate where the requisite number of replicates on all units cannot be included in a single run due to time constraints or instrumental constraints. It is also useful where a measurement process is prone to drift towards the end of longer runs; in that circumstance, better precision (measured as the within-unit standard deviation) can be obtained using several short runs in a randomized block design than can be achieved in a single run.

NOTE 1 In a classical statistical analysis (using ANOVA calculated from sums of squares), the degrees of freedom for the between-unit standard deviation remains equal to a–1 for a simple randomized block design, where a is the number of units. However, the residual degrees of freedom, which is associated with the within-unit variance, is reduced by one per additional run.

NOTE 2 In a simple randomized block design as described above, the calculated within-unit standard deviation includes both variation due to the measurement procedure and variation due to any within-unit heterogeneity.

7.6.4 Balanced nested design

In a balanced nested design for a units, a number of runs n_r are conducted, each including the same number a_w of units, which are each measured with n_0 replicates, all replication being carried out within the same run. Each run is randomized separately. Figure 4 c) shows an illustrative example.

Nested designs reduce the degrees of freedom associated with the between-unit standard deviation. For these designs, the number of units studied should be increased to retain sufficient degrees of freedom. This usually requires adding a minimum of one additional unit for each additional measurement run in a nested design. In addition, if the statistical method to be used relies on a balanced design, sufficient further RM units should be added to ensure that every measurement run includes the same number of RM units.

NOTE Balanced designs simplify analysis using classical ANOVA based on sums of squares. This makes it advantageous to maintain the same number of units per group if analysis of variance is to be performed manually or by software (such as spreadsheet applications that provide basic ANOVA functionality). More recent variance estimation methods (see 7.7.6) are computationally more complex but much less affected by unbalanced designs in which different numbers of RM units or measurement replicates appear in each group.

EXAMPLE An RM producer using a measurement procedure capable of obtaining up to 12 observations in a single run prepares 2 000 units of a candidate RM. Formula (1) yields $N_{\min}$ = 12,6, or 13 units, if the study is completed in a single run. Because the RM producer cannot obtain duplicate observations on 13 units in a single run, the RM producer elects to use a nested design with three runs. This involves adding two additional runs, and the number of units is accordingly increased from 13 to 15. The RM producer therefore measures 15 units in three groups of five units.

7.6.5 Alternative strategies

It is not always possible to carry out one of the designs listed above. For example, reference material units might not permit subdivision for replicate analysis, perhaps because the amount of material provided is sufficient only for a single measurement, because units are insufficiently stable once opened, or because the units are indivisible (for example, individual Charpy test pieces). Other situations include very short measurement runs (for example, three observations per instrument run) or a need to allow for a further source of variation in addition to run effects.

Where units cannot be subdivided, the most common strategy is to measure each unit once and to compare the standard deviation (or, more commonly, variance) of those measurements against an independent estimate of the measurement repeatability. For example, the variance of single measurements on 10 units might be compared with the variance of 10 observations on a pooled and carefully homogenized preparation of the same material, or to a good estimate of precision based on method validation studies. The difference in variance is then taken as an estimate of the between-unit variance.

Where the available run length is too short to permit a complete randomized block or nested design, strategies may include, for example:

- random allocation of replicates among short runs, followed by statistical analysis as for the basic design. This can increase the calculated within-unit variance because the run effect is then included in the within-group term;
- use of 'balanced incomplete block' designs in which a carefully chosen set of replicates is obtained in each of many short runs. Balanced incomplete block designs are beyond the scope of this document but are described in reference texts on experimental design[11]. Statistical analysis may use classical or other methods.

Where multiple sources of variation are to be controlled, other available designs may be used, including, for example, replicated Latin square designs[12] or other structured experimental designs[11].

7.7 Evaluating a homogeneity study

7.7.1 Initial inspection for measurement trends and outliers

7.7.1.1 Measurement trends

The experimental data (including any quality control observations) should first be inspected for trends in the measurement process. This is most effectively done by visual inspection of a plot of the data in experimental run order. Regression analysis, such as linear regression using a simple straight-line model, may be used to test for a statistically significant trend if an approximately linear trend is apparent.

A measurement trend in a properly randomized homogeneity experiment will generally inflate the within-unit variance, potentially obscuring important between-unit effects. If a significant trend in the measurement results is confirmed, action should be taken to minimize its effects on subsequent analysis. Typical actions include:

- Reject the affected measurement run(s) and repeat the measurements after eliminating the source of the trend.
- Apply a correction to the data based on a suitable smoothing function (for example, a linear regression) before conducting subsequent analysis of variance. Prior correction for a trend reduces the residual degrees of freedom and this should be taken into account in subsequent variance estimation.
- Include the trend in the data analysis. This procedure adjusts properly for any loss in degrees of freedom, though special statistical software and appropriate statistical expertise are required.

NOTE 1 Examples of these options can be found in Annex C, example C.4.

NOTE 2 The RM producer is expected to define criteria to determine whether a trend is significant. In the absence of other criteria, statistical significance at the 95 % level of confidence is usually taken as an indication of a significant measurement trend in homogeneity studies.

NOTE 3 In some cases, for example in a nested or randomized block design with three or more runs, omitting a single failed run can, if the precision in remaining runs is sufficiently good, leave sufficient data for effective homogeneity assessment without repeating the failed run.

7.7.1.2 Outliers in the basic homogeneity study

Many measurement results are generated as part of a homogeneity study and technical issues can result in occasional outlying results due to instrumental or other attributable causes. The cause of outliers should be investigated and appropriate corrective action should be applied. Corrective action in such a case may include, for example:

- removal of a single outlying observation or of all results for a unit of material which shows one such observation (see NOTE 2 below);
- re-measurement of a suspect unit or subsample within a short period of time;
- repeating a complete measurement run or a complete experiment affected by several outliers.

It is not appropriate to reject homogeneity study results for a reference material unit on the basis of an extreme mean value for that RM unit.

NOTE 1 Outliers can be identified by visual inspection or by using outlier tests such as Grubbs tests. ISO 5725-2[75] describes Grubbs tests. ISO 16269-4[76] describes graphical outlier tests and other tests for multiple outliers.

NOTE 2 Within-unit outliers can be caused by within-unit heterogeneity. If within-unit heterogeneity is likely, it is prudent to investigate further before discarding any outlying replicate result.

NOTE 3 Omitting a single outlying value from a single unit will result in imbalance. Where the statistical method in use relies on a balanced design, the data for that unit can be omitted provided that the non-outlying values for that unit are consistent with the remainder of the data set and that the number of units remaining in the study is sufficient.

NOTE 4 Larger homogeneity studies (for example, studies involving 20 or more RM units for a given property) can naturally incur more than one within-unit outlier due to instrumental effects. After investigation into the cause of the outlying values, it can be appropriate to discard more than one such observation.

7.7.2 Inspection for processing trends

Plotting unit means in processing order provides a simple method of checking for processing run effects. Alternative methods include plots of residuals against unit number after fitting a model that includes measurement run effects; this can be particularly useful after a balanced nested design as unit means can be influenced by measurement run differences. It is not appropriate to apply mathematical corrections to remove trends arising from the processing of the reference material. If such a trend is detected and the material is still found to be useful, the uncertainty of homogeneity should be estimated in such a way that it covers the observed trend as well as any random between-unit variation around this trend.

7.7.3 Evaluation of the between-unit term – basic design

The basic design described in 7.6.2 may be analysed by one-way analysis of variance as described in B.1. The calculation assumes there are a RM units and n_i measurements on RM unit i, where i = 1, 2, ..., a. Usually, simple homogeneity studies are intended to be balanced designs with the same number n_0 of replicate measurements for all RM units, so that $n_1 = n_2 \ldots = n_a = n_0$. If a one-way analysis of variance approach is then used as described in B.1, the between-unit standard deviation s_{bb} can be computed from the between-group mean square M_{between}, the within-group mean square M_{within} and the number of replicates per unit n_0 using (with symbols as in Annex B).

$$s_{\text{bb}}^2 = \max\left(\frac{M_{\text{between}} - M_{\text{within}}}{n_0}, 0\right) \tag{3}$$

In this simple case, the between-unit variance $s_{\text{bb}}{}^2$ is identical to the (squared) between-unit homogeneity contribution to the uncertainty, u_{bb}^2 (see 7.11 [16]).

NOTE 1 Annex C, Example C.1 describes an example of a homogeneity study using one-way analysis of variance.

NOTE 2 It can often be assumed that the true RM unit means are normally distributed, though this assumption is not necessary for the calculation given at Formula (3). Likewise, it can often be assumed that the within-unit random measurement error is normally distributed[13].

NOTE 3 If the condition stated in Formula (2) in 7.5.1 is not met, see 7.8.

7.7.4 Evaluation of the between-unit term – randomized block design

A randomized block design involving a RM units with n_0 replicates, with one replicate per unit per run and n_{r} runs, may be analysed using two-way analysis of variance without replication. The statistical model is

$$x_{ij} = \mu + A_i + B_j + \varepsilon_{ij} \tag{4}$$

where μ is the population mean, A_i the effect of unit i and B_j the effect of run j on the result. This is identical to the model for the basic design (B.1) except for the addition of B_j, the effect of run j on the results. The determination of the between-unit variance s_{bb}^2 is not dependent on the distribution of A or B, though large run effects could in principle affect the residual term, indirectly, if the precision depends on mean observed level. Usually, however, run effects that are large enough to show such an effect would invalidate the study entirely. In addition, the analysis assumes that run effects and unit effects are independent and additive. This is also a reasonable approximation for the relatively small effects (compared with the within- and between-group standard deviation) expected in homogeneity studies.

Analysis of variance applied to a randomized block design, following the model at Formula (4) and with one observation per unit per run, leads to a between-run mean square M_{B} together with, as before, a between-unit mean square M_{A}, and a residual mean square M_{within}.

The analysis of variance table also includes associated degrees of freedom for each term. The residual mean square M_{within} is an unbiased estimate of the repeatability variance s_r^2, although inspection of the ANOVA table will show that the number of degrees of freedom is $(a{-}1)(n_0{-}1)$ rather than $a(n_0{-}1)$ as in the basic design. The between-unit standard deviation s_{bb} is calculated as in Formula (5) and has $(a{-}1)$ degrees of freedom.

$$s_{\text{bb}}^2 = \max\left(\frac{M_{\text{A}} - M_{\text{within}}}{n_0}\right) \tag{5}$$

NOTE An example of treatment of a randomized block design can be found in Annex C, Example C.3.

7.7.5 Evaluation of the between-unit term – balanced nested design

Variance components for balanced nested designs using a units equally distributed among n_{r} runs and with m_{w} units per run may be calculated using analysis of variance methods described in ISO 5725-3[14], which includes nested designs, or by following standard statistical texts such as Reference [15]. The statistical model is given by Formula (6):

$$x_{ijk} = \mu + A_{ij} + B_j + \varepsilon_{ijk} \qquad (6)$$

where x_{ijk} is the kth observation for unit i ($i = 1, 2, ..., m_{\mathrm{W}}$) in run j ($j = 1, 2, ..., n_{\mathrm{r}}$), A_{ij} is the effect of unit i in run j, B_j is the effect of run j and ε_{ijk} ($k = 1, 2, ..., n_0$) the usual residual error term. Analysis of variance for a balanced nested design with units nested in runs will again generate a between-run mean square M_{B} together with a between-unit mean square M_{A}, and a residual mean square M_{W}. Assuming units nested within runs and with n_0 replicates per unit, the between-unit term is again estimated using Formula (3), though with $a-n_{\mathrm{r}}$ degrees of freedom instead of $a-1$ as in the basic design.

Two way ANOVA for the nested design is described in B.2. For analysing a homogeneity study using a units equally distributed among n_{r} runs and with m_{W} observations per RM unit, a, n_{r} and m_{W}, respectively, replace p, all b_i and all n_{ij} in B.2.

7.7.6 Other homogeneity designs and alternative estimation methods

Other designs will normally require analysis using specialized statistical software capable of variance component analysis for a wide range of designs. Some such software continues to use variance estimation based on computing mean squares and equating the mean squares to expected mean squares; this is the method used in 7.7.3 to 7.7.5. Other statistical methods have become available, however; of most interest are maximum likelihood methods[15] and Bayesian analysis[77]. Although detailed calculations are beyond the scope of this document, some guidance is given below for the selection of such methods.

Maximum likelihood methods are among the most general and widely used methods of variance component estimation. The recommended maximum likelihood approach for estimating variances in reference material work is the variant known as Restricted (or, sometimes, 'residual') Maximum Likelihood estimation (REML). This provides essentially identical results to those of analysis of variance for balanced designs with no missing data, but also deals well with complex designs, missing data and appreciable imbalance. REML may therefore be used in place of the ANOVA-based methods above as well as in a wider range of circumstances.

Bayesian analysis is increasing in popularity as advanced computational methods become available. Bayesian analysis, however, depends on the use of a prior probability distribution for estimated parameters, including variance parameters. It also generally requires careful specification of the statistical model and in general is more computationally intensive. Although Bayesian methods give good results in experienced hands, they can also give unexpected and misleading results when the model is incorrectly specified or inappropriate priors are chosen, especially for the comparatively small data sets common in reference material production. Bayesian methods for variance component estimation in reference material certification should therefore be used only with the assistance of a qualified statistician experienced in Bayesian analysis of small data sets.

7.8 Insufficient repeatability of the measurement procedure

It is not always feasible to perform a homogeneity study with a measurement procedure that is sufficiently repeatable (that is, that meets the requirement of 7.5). In those cases, an alternative approach may be used that estimates an upper limit for the between-unit term.

Discussions of various approaches to obtain an uncertainty estimate that accounts for insufficient repeatability of the measurement procedure are given in References [16] and [78]. An example of application of the recommendations of Reference [16] is given as Annex C, Example C.2.

7.9 Within-unit homogeneity

7.9.1 Assessing the need for within-unit homogeneity study

Within-unit homogeneity is an issue that arises when the minimum sample size specified in the instructions for use is smaller than the size of a complete RM unit. It is always important to ensure that the specified sample intake is sufficient, and a between-unit homogeneity study alone does not provide such assurance. Where the minimum sample size is small and there is a high risk of appreciable within-unit heterogeneity, an experimental within-unit homogeneity study should be carried out. Situations involving a high risk of appreciable within-unit heterogeneity include one or more of the following:

— minimum sample size substantially less than the unit size;

— a material prepared by mixing powders or granular materials;

— a material prepared by mixing a small quantity of one component into a bulk matrix; or

— a material with previously known within-unit heterogeneity.

Experimental assessment may consist of a test for significant within-unit heterogeneity or a determination of minimum sample size.

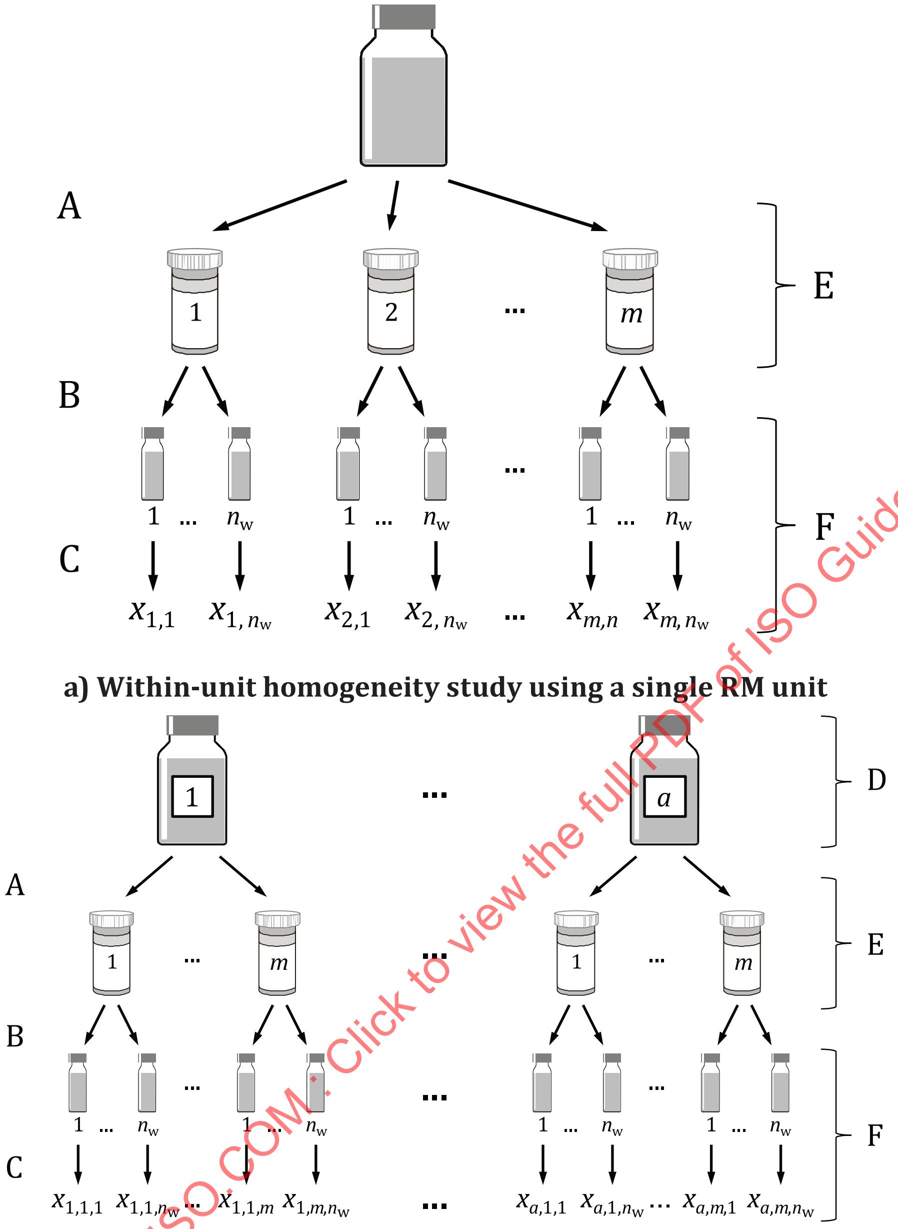
7.9.2 Testing for significant within-unit heterogeneity

A test for significant within-unit heterogeneity uses a design similar to between-unit homogeneity studies, with the exception that the variance of interest is the within-unit, between-subsample term. Figure 5 shows some possible designs for a within-unit homogeneity study. In the simplest [Figure 5 a)], m test portions, in the size range of the envisaged minimum sample size, are drawn from a single RM unit and each is prepared and measured $n_W \geq 1$ times. Where insufficient aliquots can be drawn from a single RM unit, or in order to obtain more representative information, more RM units can be included, typically with reduced within-unit replication [Figure 5 b)].

A test for within-unit heterogeneity should provide at least five degrees of freedom for the within-unit term. This can be achieved by, for example, taking six or more subsamples from a single unit [$m \geq 6$ in Figure 5 a)]; taking three subsamples from each of three or more units [$a \geq 3$, $m = 3$ in Figure 5 b)]; or by taking duplicate subsamples from each of five or more units [$a \geq 5$, $m = 2$ in Figure 5 b)]. Where subsamples are taken from multiple units, the analysis should include a between-unit term in a manner analogous to including a run term in the balanced nested design at 7.7.5.

Where replicate analyses can be performed on each subsample, analysis of variance followed by an F-test for significant between-subsample variance may be applied. For a single RM unit with aliquots measured multiple times, one-way ANOVA with subsample as the grouping factor can be used as in B.1; where multiple units are subsampled and multiple aliquots taken, as in Figure 5 b), two-way ANOVA for nested designs is appropriate (B.2). Where only a single observation can be made per subsample ($n_W = 1$), the variance of the observations (or, from multiple units, the within-group term from one-way analysis of variance) should be compared with a reliable estimate of the repeatability of the procedure, usually using an F-test for significantly greater variance for results on different subsamples. For the case where the repeatability is well established with high degrees of freedom, such an assessment can be done by comparing the observed variance with the known repeatability variance using a χ^2-test[1].

NOTE Even where it is possible to perform replicate measurements on a subsample, it is not always possible to undertake replicate preparation of a subsample as indicated in Figure 5 b), particularly where the subsample size has been chosen to match the proposed minimum sample size. This results in a design intermediate between those shown in Figure 5 a) and in Figure 5 b).


A
B
C
E
F
1
2
...
m
1 ... n_w
$x_{1,1}$ x_{1,n_w} $x_{2,1}$ x_{2,n_w} ... $x_{m,n}$ x_{m,n_w}
D
a
$x_{1,1,1}$ $x_{1,1,n_w}$ ... $x_{1,1,m}$ x_{1,m,n_w}
$x_{a,1,1}$ $x_{a,1,n_w}$ ... $x_{a,m,1}$ x_{a,m,n_w}

a) Within-unit homogeneity study using a single RM unit

b) Within-unit homogeneity study using multiple RM units

Key

A subsampling
B preparation
C measurement
D contributes to the observed between-unit variation
E operations contributing to observed within-unit, between-subsample variation
F operations contributing to observed within-subsample variation

Figure 5 — Schematic designs for within-unit homogeneity study

7.9.3 Assessing minimum sample size

7.9.3.1 Experimental determination

The minimum sample size may be determined experimentally by carrying out a within-unit homogeneity study for different sizes of sample intake. As the within-unit homogeneity standard deviation depends on the number of particles carrying a certain property, the observed standard deviation is expected to increase as the sample intake decreases and to converge to the measurement procedure repeatability as the sample intake increases. This can form the basis for a sample intake study, such as the following procedure:

a) determine the observed standard deviation of replicates at a range of different sample intakes (mass or volume of subsamples);

b) plot the observed standard deviation against the sample intake;

c) either

  - use the plot to demonstrate that the sample intake has no effect on the observed standard deviation in the range assessed, or

  - choose a sample intake at which the estimated standard deviation is indistinguishable from the repeatability standard deviation for the measurement procedure.

Since the number of degrees of freedom is small, it is important that the range of sample intakes studied be sufficient to yield a significant difference if there is appreciable within-unit heterogeneity. Sample intakes covering a range of at least a factor of five (for example, from 10 mg to 50 mg) should therefore be included in such an experiment, where possible.

Where it can be shown that the between-subsample variance is inversely proportional to the mass taken, Ingamell's sampling theory applies[17]. In this circumstance, the mass m_{I} required to achieve a particular between-subsample relative standard deviation s′ is given by Formula (7):

$$m_{\mathrm{I}} = K_{\mathrm{I}} \big/ (s')^2 \qquad (7)$$

where K_{I} is Ingamell's sampling constant.

NOTE Ingamell's sampling constant is the mass of material that leads to a between-subsample relative standard deviation of 0,01 (1 %) and can be determined from the between-subsample relative standard deviation at other subsample masses.

7.9.3.2 Other methods

Minimum sample size may be set using other data or by experience. Other data that may inform the choice of minimum sample size includes, for example:

- sample intake used for obtaining results with acceptable precision in a stability study;
- sample intake providing acceptable precision in a characterization study carried out in a single laboratory;
- sample intake from laboratories achieving acceptable precision in an interlaboratory characterization;
- sample intake used for the homogeneity study.

7.10 Check for sufficient homogeneity

Having determined the between-unit and (where appropriate) within-unit standard deviation, the RM producer should confirm that the variation within and between units is sufficiently small for the intended use of the material. Such confirmation may include, for example:

a) comparison of the between-unit and/or within-unit standard deviation to the uncertainty associated with characterization to confirm that the standard deviation(s) are small compared with the characterization uncertainty (for example, $\mathrm{s}_{\mathrm{bb}} < u_{char}/3$);

b) calculation of the combined uncertainty u_{CRM} of the certified value, with allowance for heterogeneity, and confirmation that the uncertainty of the certified value is acceptable for the intended use;

c) for reference materials where the combined uncertainty u_{CRM} will not be calculated, checking that the between-unit standard deviation s_{bb} is small compared with the typical interlaboratory reproducibility standard deviation s_{R} in the field of use. Ideally s_{bb} should be less than $s_{\mathrm{R}}/3$;

d) confirmation, for example by use of an F test, that the between-unit term is not statistically significant at the 95 % level of confidence.

NOTE 1 Lack of statistical significance does not prove the absence of heterogeneity, but can be a useful practical criterion when it is not intended to provide an uncertainty with a property value and the procedure used for the homogeneity study is typical of procedures employed by end users of the material.

NOTE 2 A method of confirming that the between-unit term is not significantly greater than a predetermined upper limit is given in ISO 13528[10].

7.11 Uncertainty evaluation from homogeneity studies

The uncertainty for a certified value (see 10.2) should include an allowance u_{hom} for any detected heterogeneity. u_{hom} should not be less than the between-unit standard deviation s_{bb} determined from experimental homogeneity studies. s_{bb} can be zero or negligible compared with other uncertainties, in which case the term u_{hom} may be omitted unless within-unit heterogeneity is important.

Where there is significant within-unit heterogeneity at the minimum sample size to be recommended in the instructions for use, the uncertainty associated with a certified value should be increased by the inclusion of a further allowance u_{wb} that should not be less than the between-subsample term derived from experimental studies (see 7.9, and Notes to B.2 for estimation of u_{wb}). The uncertainty u_{hom} associated with heterogeneity is then given by Formula (8):

$$u_{\mathrm{hom}} = \sqrt{u_{\mathrm{bb}}^2 + u_{\mathrm{wb}}^2} \tag{8}$$

The contribution u_{hom} to the uncertainty of a certified value may also be increased to allow for insufficient precision of the measurement procedure used for homogeneity studies, as provided for in 7.8.

8 Assessment and monitoring of stability

8.1 Preamble

Stability of the properties of interest is one of the key features of all RMs. It is important that the value for each property of interest is, at the time of use, consistent with the stated value on the certificate or other documentation accompanying the material.

The value of each property can change over time for a variety of reasons, to different degrees, and at different rates depending on the conditions. Three sets of conditions are particularly important: conditions during long-term storage at the RM producer’s facilities, conditions during transport to the user’s premises, and the specified conditions of storage and use at the user’s premises.

The form and rate of change can differ considerably for different materials. Some change little or not at all under a wide range of conditions. Some change rapidly under ambient conditions and require low temperature storage. The form of degradation can differ: some materials change almost linearly over long periods; some can undergo autocatalytic or less predictable rapid change after a period of stability. Some can change rapidly during an initial period after processing and then remain stable over long periods. For some, change can affect some units of the material much more than others, leading to changes in between-unit homogeneity rather than a consistent change in value for all units.

These very different patterns of change, under different conditions, can be hard to predict even after extended experimental study. There is, therefore, always some risk that the value of one or more properties will change unexpectedly during the life of a reference material, or even during the life of a particular batch of a reference material that is produced regularly. Accordingly, the RM producer is expected to manage material processing, storage, packaging, transport conditions, post-certification monitoring and advice to end users so that the risk of unexpected change is as small as reasonably possible.

The RM producer should:

a) assess, by experimentation if necessary, the stability of all relevant properties of a reference material under proposed storage conditions and choose pre-treatment, packaging and storage conditions accordingly;

b) assess, by experimentation if necessary, the stability of all relevant properties of a reference material under planned conditions of transport, and choose transport conditions to maintain stability during transport;

c) establish any necessary advice on storage and use of the material to maintain stability at the user‘s premises;

d) select a scheme for monitoring the stability of materials held in long-term storage that permits prompt detection of change, taking into account the possible rate of change;

e) where the stability of a certified value cannot be ensured, make due allowance in the stated uncertainty for possible change in the value prior to use or, where the change with time can be predicted, provide a means of correcting the certified value and its uncertainty for expected change over time;

f) where repeated sampling from a reference material unit, or repeated use of an entire reference material unit, is permitted by the instructions for use, assess the possible effects on the stability of the material and take appropriate action.

The following subclauses provide guidance on the need for experimental study and, in the common case where experimental studies are required, provide guidance on their nature and conduct. Guidance is also provided on the necessity for, and assessment of, uncertainties associated with possible instability for those cases where there remains an appreciable risk of steady change over time.

NOTE 1 Conformance with a) to f) above is a requirement of ISO 17034.

NOTE 2 It is expected that for most materials, RM producers will select a combination of pre-treatment, packaging, storage, transport conditions and a monitoring scheme that lead to a reasonable expectation of negligible change over time.

NOTE 3 The results of monitoring can additionally inform future stability management of the same, or related, RMs.

NOTE 4 Any means of correcting assigned values for predicted change over time will, for certified values, additionally require the estimation of the uncertainty associated with the corrected value.

NOTE 5 A material can be stable with respect to one property, even though other property values can change appreciably when measured over time.

NOTE 6 Most statistical treatments for data describing change over time use a mathematical model to estimate continuous change. However, while a model can be checked for adequacy across the time course of a given study, the very wide range of forms for possible future degradation prevent reliable prediction. Nor is it practical to include uncertainty components accounting for possible alternative degradation models; the large divergence in behaviour of different models would render such uncertainties impractically large. In practice, the only reliable way of estimating change over extended periods at a given set of conditions is to observe the material over the complete period of interest. For this reason, whatever model is used to test for change during the study and to predict future change, stability monitoring remains essential for any material intended to remain available for extended periods.

8.2 Assessment of stability

8.2.1 Requirement for stability assessment

The stability of all RMs should be assessed. The assessment of stability may involve an experimental study to estimate the remaining degree of instability of the candidate RM after processing or to confirm the stability of the material. Any detectable changes in the property values due to long-term instability or the effect of transportation should be considered in the claims for the property values, and the uncertainty for these estimates of change should be included in the uncertainty for certified values.

Where a reference material is produced in repeated batches that are not individually tested for stability, the RM producer should additionally assess the risk of changes in stability from one batch to the next and should verify the stability of a sufficient number of different batches experimentally to provide confidence in the stability of all batches (see 8.2.4.2).

8.2.2 Types of (in)stability

Two types of (in)stability are particularly relevant in the production of reference materials:

- the long-term stability of the material (i.e. the stability of the material during the period of validity under specified storage conditions);
- the stability under reasonably expected conditions of transport ("transportation stability").

The long-term stability of a reference material is associated with the behaviour of the RM on the shelves of the RM producer or when stored according to prescribed conditions by the user. Long-term stability of the material should be assessed prior to distribution of the material to users.

The transportation 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 certified values. The RM producer should, prior to distribution of the material to users, assess the stability of each property of interest under reasonably expected conditions of transport.

NOTE 1 Transportation stability is often referred to as "short-term stability".

NOTE 2 It is often useful to know what might happen to the sample if the intended transport conditions are not maintained. This knowledge allows the RM producer to give better advice and, from the perspective of the user, a better product. Stability assessment can therefore be conducted to assess instability under more extreme conditions than those expected on the basis of planned storage and transport conditions.

8.2.3 General methods for assessment of stability

The assessment of both long-term and transportation stability should include one or more of the following:

- consideration of all the physical, chemical and biological properties of the material that might reasonably affect stability, including the particular chemical or biological species certified;
- review of published literature data on stability of related materials;

- review of stability assessment or monitoring data on related materials;
- planned experiments, whether real-time or accelerated stability studies;
- experiments to test the effect of different storage arrangements, including container integrity and stabilization or preservation methods.

Where stabilization is used, the RM producer should define the method of stabilization and verify (by any of the methods above) that the stabilization method used is appropriate for the use of the material and effective[18].

Stability assessment should consider potential effects of re-use or repeated subsampling (including, for example, the effects of reopening, re-freezing or humidity) when this is permitted under stated conditions for use. Where it is not feasible to include the effect of repeated subsampling or re-use in the stability assessment, the instructions for use should not permit re-sampling or re-use.

NOTE 1 “Related” materials above can include appreciably different materials from the candidate reference material that share some, but not all, relevant characteristics. For example, the same certified property in a different matrix or solvent, the same matrix type but different certified properties, etc.

NOTE 2 8.2.4 and 8.9 include further provisions related to repeated subsampling or repeated use.

8.2.4 Need for experimental study of stability

8.2.4.1 General requirements on the need for experimental study of stability

For reference materials for which the RM producer has little or no prior information on stability under the planned storage and/or transport conditions, or where the effects of permitted re-sampling or re-use are not known, an experimental assessment of stability should be conducted.

Experimental studies are not necessary if the RM producer has prior information on stability from closely similar materials held for an extended period under the same planned storage conditions.

8.3 to 8.8 give guidance on the form and conduct of experimental stability studies.

NOTE “Closely similar” materials are materials characterized for the same properties, which share the same matrix composition, processing conditions, similar or less effective packaging, etc.

8.2.4.2 Need for experimental study of stability in repeated batch production

The RM producer should normally verify the stability of at least two batches experimentally, in addition to the batch chosen for experimental characterization of stability, when

- an experimental study is considered necessary,
- the material in question is to be produced in repeated batches, and
- the RM producer intends to use the stability data from one batch to represent the stability of the following batches.

NOTE Verification can be a simple test to confirm that different batches behave similarly or, for successive batches, do not change over their lifetime, while the experimental assessment of stability typically involves an extended study aimed at determining rates of change.

8.3 Classification of stability studies

8.3.1 General

Approaches to a stability study can vary significantly. The simplest approach is to take two independent measurements at different points in time, in order to draw conclusions about future stability based on change over the elapsed time. Other approaches include multi-point approaches, which enable the

RM producer to model more complex stability behaviour and/or draw conclusions about change over shorter periods than the time elapsed between the first and last measurement in the stability study.

Stability studies can be classified into

- classical and isochronous studies, according to the conditions of measurement,
- real-time studies and accelerated studies, according to the conditions and treatments applied, and
- transportation and long-term stability studies, according to whether they are aimed at evaluating stability under transport conditions or long-term stability in storage.

In addition, it can be necessary to examine particular conditions, such as packaging features or humidity, and this can require special experimental designs.

The following subclauses provide guidance on these different design options.

8.3.2 Classification according to conditions of measurement

8.3.2.1 Classical stability studies – Intermediate conditions of measurement

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 (e.g. one sample immediately, one after three months, the next one after six months, etc.). This design, in which the measurements are carried out under intermediate conditions of measurement (sometimes called within-laboratory reproducibility conditions), can lead to a relatively high uncertainty when instability of the measurement system contributes significantly to the dispersion of the measurement results.

8.3.2.2 Isochronous stability studies – Repeatability conditions of measurement

Isochronous designs use storage under reference conditions to allow RM units exposed to different degradation conditions to be measured in a short period of time, ideally under repeatability conditions[19]. Reference conditions are a set of conditions under which the properties of interest can be reliably expected to be stable, or can be a chosen baseline level. The word “isochronous” emphasizes that the measurements are made at the same time, rather than distributed over the time span of the stability study, as is the case in the classical approach. This use of repeatability conditions is expected to lead to improved measurement precision over the course of the study, thus improving the power of the stability study. As a consequence, the isochronous stability study, in theory, leads to a smaller uncertainty than that of the classical study, depending on the difference between the repeatability and the (within-laboratory) reproducibility of the measurements. A prerequisite for this design 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.

A more detailed description of isochronous studies is given in 8.5.3.

8.3.3 Classification according to stability study duration and conditions

8.3.3.1 Real-time stability studies

In a real time stability study, the stability of a material is studied under the storage or transport conditions that are intended for the RM. This means that one week/month/year of the stability study gives information on the behaviour of the material over a one week/month/year period. This type of study has the advantage that it does not require any assumptions about the effects of different conditions on the stability, because the conditions used in the experimental study are the same as those intended for use in transport or storage.

8.3.3.2 Accelerated stability studies

In an accelerated study, multiple experiments are performed at conditions that are more extreme than the storage or transport conditions intended for the RM and that aim at inducing more rapid degradation than would be experienced under the intended storage conditions. The degradation rate at the conditions of interest is then estimated, for example, by regression analysis over the various experimental conditions. The most frequent example is testing at several temperatures and estimating a degradation rate via the Arrhenius equation. This type of study has the advantages of reducing the total time required and of increased confidence from the use of information from more extreme exposure conditions. The major disadvantage is that the degradation mechanism or its rate-determining step can change with different conditions under study, particularly the temperature.

NOTE 1 Both real-time and accelerated studies can be organized as classical or as isochronous studies.

NOTE 2 The use of accelerated studies can provide confidence in stability for periods substantially longer than the study duration and is particularly useful where early availability of the material is required.

NOTE 3 Light, moisture and temperature are common examples of factors that can accelerate degradation.

8.3.4 Classification by study objective

8.3.4.1 Transportation or other short-term stability studies

Transportation stability is a property of the material referring to stability under expected transport conditions. For this, the behaviour of the material and its property values are studied under (as a minimum) the intended conditions for packaging and transport. The more restrictive the intended conditions for transport are, the smaller is the range of conditions that should be included in the transportation stability study.

When no previous experience is available concerning a particular type of material, a transportation stability study should be carried out to gain information concerning the appropriate conditions for transport. The duration of and conditions included in a transportation stability study should reflect the duration and conditions reasonably expected in transporting a unit of the RM to the user's premises. Unless there is reason to the contrary, these conditions should include extreme temperatures that might reasonably occur during international transport for a period that is at least as long as that allowed for transport of the RM. For example, if the proposed transport time is restricted to 3 weeks, a short-term stability study of 3 to 4 weeks will suffice.

If experience of the stability of similar materials is available and indicates that closely controlled transport conditions are required (for example, transportation packed with dry ice or ice-packs), the study may be restricted to the intended conditions of transport.

Other short-term studies can be necessary, for example to establish the suitability of expected storage conditions at the end user's premises. Studies to establish the suitability of alternative storage for short periods at the end user's premises should reflect the conditions permitted by the instructions for use (see 8.9). It is not usually necessary to explore more stringent conditions for this purpose.

NOTE It is assumed in this document that transport conditions will normally be sufficiently closely controlled that the effect of transport does not significantly affect the property values and does not require an increase in uncertainty for certified values. Transportation stability studies are therefore assumed to be designed primarily to check for significant change rather than to provide an accurate quantitative estimate of change.

8.3.4.2 Long-term stability studies

Long-term stability studies are conducted to assess stability under storage conditions specified for the lifetime of the product. Real-time long-term studies typically last 12 months or more; accelerated studies are typically shorter but include more extreme conditions. The period of validity of the certificate is also ensured by stability monitoring after release. In general, the fewer data that are available about the stability of a property value in a material, the more extensive the long-term stability and post-certification monitoring should be. Where time-to-market for new materials is crucial, it is possible to limit the long-term stability study to less than 12 months and perform frequent monitoring to complement the limited data available before certification.

8.3.5 Designs for different storage and treatment conditions

It is sometimes necessary to examine the effects of more than one storage or treatment condition. For example, it can be necessary to assess effects of high humidity, high and low temperature, different stabilization methods, effect of exposure to light, or different packaging options.

Experiments for this purpose may include:

- individual study of each possible effect on stability;
- multi-factor experiments intended to evaluate the effect of several storage and/or packaging treatments simultaneously;
- 'worst case' studies that expose a material to the most extreme combination of conditions, on the basis that stability under the most extreme conditions provides evidence of stability under less extreme sets of conditions.

Where multiple storage and/or packaging conditions need to be studied, the use of factorial or fractional factorial designs is encouraged for efficiency reasons. Where used, such experiments should allow for the possibility of interaction between different factors. Two-factor interactions, such as the combined effect of humidity and temperature, are often important and the experimental design should allow for estimation or testing of these, where practicable.

NOTE 1 Guidance on factorial and fractional factorial experimental designs can be found in Reference [11] and their use for stability studies (of drug substances and products) is described in Reference [20].

NOTE 2 Where an effect is found in a multi-factor experiment, it is important to verify the effect by further study.

NOTE 3 In the experimental design literature, fractional factorial designs that allow for the estimation of main effects without bias from two-factor interactions are sometimes known as 'resolution IV' designs. Resolution V designs additionally allow separation of two-factor interactions from other two-factor interactions.

8.4 General requirements for effective stability studies

8.4.1 Overview of requirements

To obtain reliable results in a stability study, it is important to

- select a representative subset of material,
- choose a suitable measurement procedure with sufficient precision and selectivity,
- make the measurements under suitable conditions following an appropriate experimental design, and
- conduct the statistical analysis using valid statistical methods.

8.4.2 to 8.4.4 provide guidance on the first three of these points. Statistical analysis is considered in detail in 8.5.

8.4.2 Selection of units

Units for a stability study on prepared and packaged materials should normally be selected randomly from the set of packaged units.

Preliminary stability studies may, however, be conducted on smaller batches of material in order to establish appropriate packaging or pre-treatment. In such cases, it is sufficient to prepare units specifically for preliminary studies and to study all, or a subset, of those so prepared.

NOTE 1 Stability study interpretation is generally less dependent upon random selection of units than are homogeneity studies or value assignment; useful indicative information about stability is likely to be obtained from simple sampling strategies such as systematic sampling. It is, however, important to avoid any sampling scheme which might result in unrepresentative stability information. For example, selection of units from the outer regions of a large set of RM units – such as the top or side of a large container or storage area – can result in selection of material exposed to light or more extreme temperature variation. Such units might have already undergone some change and can behave unrepresentatively in further stability assessments, particularly if the material is susceptible to rapid early change and shows subsequent approximate stability.

NOTE 2 7.4.3 gives examples of appropriate sampling schemes for selecting RM units for study.

8.4.3 Suitable measurement procedure(s) for stability studies

Irrespective of the study design, the outcome is only meaningful if the standard deviation of measurement results over the study time scale, possibly in conjunction with the between-unit homogeneity, is sufficiently small.

For isochronous studies (see 8.3.2.2 and 8.5.3) or other designs which rely primarily on good repeatability, the measurement procedure(s) used should be selected primarily for good repeatability. For studies which might be affected by run-to-run variations in measurement, such as the simple classical design at a single storage condition, measurement procedures should be selected primarily for good intermediate precision.

8.4.4 Appropriate experimental design

A range of designs for the experimental study of stability is given in 8.3. Alternative designs may be used where their effectiveness can be demonstrated.

The degree of replication required in a stability study and the duration of the study depend on a variety of factors. In particular, uncertainties in predicted degradation depend both on the study duration and the number of replicates. In deciding the number of replicates, the number of different exposure times, and the number and range of different sets of exposure conditions, consideration should be given to:

- the precision available from the measurement procedure chosen for the study; very precise measurement procedures require low levels of replication;
- the need to allow for failures in individual observations or RM units. Allowing a minimum of two RM units for each combination of time and temperature (or other conditions) provides for redundancy and helps to avoid missing sets of conditions;
- the planned lifetime and monitoring frequency for the material. Long monitoring intervals require either lengthy stability studies or higher replication to provide reliable predictions of degradation;
- the need to verify linear (or other) model behaviour. To provide some check on linearity, a minimum of three observation times is essential in an isochronous study; for non-isochronous studies, where run effects may be important, a minimum of four points in time, with replication, is required;
- the need to observe the material under proposed storage conditions. Studies should include (but need not be limited to) exposure under the proposed conditions of storage;
- the need to examine interactions between factors affecting stability.

For materials with little prior information available, the number of units and replicates used should, in addition to the considerations above, be sufficient to provide small uncertainties for predicted degradation.

Where the available measurement procedures do not have sufficient precision for reliable determination of stability, additional replicates at each combination of point in time and conditions and/or additional points in time should be included. The preferred nature of replication depends on the principal sources of variation, as follows.

— Where the measurement repeatability is the principal source of variation, the number of replicated measurements on each unit and/or the number of units studied at each combination of point in time and other conditions should be increased.

— Where RM heterogeneity (represented by the between-unit standard deviation s_{bb}) is an important source of variation, the number of units studied at each time/condition combination should be increased.

— Where measurement variation over time in a classical stability study is important, the number of points in time should be increased.

NOTE 1 When between-unit heterogeneity arises from a known trend it is important to allocate RM units to time/temperature treatments so as to avoid bias. This is usually achieved by ensuring random allocation of units to treatments.

NOTE 2 Exposure times in a stability study are not necessarily equally spaced.

NOTE 3 The International Conference on Harmonization (ICH) describes some efficient designs for study of multiple factors affecting stability[20].

8.5 Evaluation of stability study results

8.5.1 General considerations for stability study data treatment

Data treatment for stability studies should take account of the particular study objective, the experimental design used, and the sources of variation that might affect the results.

For most basic stability studies, the objective is either to test for any important change over time in storage or to estimate the rate of change of property values over time. In studies that examine the effect of single or multiple storage conditions (8.3.5 above), data analysis is normally intended to provide a test for significant effects. Some designs include elements of both of these objectives.

The experimental design used will affect the data analysis options. A design restricted to two sets of conditions – for example, two extreme points in time – can (assuming replication at each point) be assessed using either a test for significant change between two groups or by regression analysis, and in this simple case the two procedures are equivalent. A design that follows a property value over time at a single set of storage conditions is typically assessed using modelling (such as linear regression, 8.5.2.3) to estimate rates of change or by analysis of variance to test for significant differences over time, but cannot be assessed by a simple two-group test. An accelerated study, which follows a material over time at several storage conditions, can require treatment using a more general model that allows estimation of different rates of change under different conditions (for example, the temperature dependence of rates of degradation).

The sources of random variation that affect the result also affect data analysis. For the simplest design, in which only a single source of random variation is present, simple linear regression as described in 8.5.2 is usually sufficient. For studies involving measurements at different times, and in which the measurement system might show run-to-run variation in addition to within-run variation, data analysis should be chosen to allow for the additional source(s) of random variation. Data analysis for such cases is considered in 8.5.5.

The following subclauses describe the analysis of stability study data. 8.5.2 describes the application of linear regression in the simplest case, and provides guidance for more complex cases including isochronous studies using one storage and one reference condition, accelerated studies, and studies with more complex error structure.

NOTE Uncertainty evaluation for certified reference materials can also use the results of stability studies.

8.5.2 The basic stability study: multiple points in time at a single storage condition

8.5.2.1 Applicability

This subclause applies to a situation where one or more measurements are taken at each of several points in time, and the random errors in each measurement are independent and share the same standard deviation and distribution (that is, errors are independent and identically distributed).

NOTE 1 The assumption of independence is not valid when the measurements are taken at different times and the intermediate precision standard deviation of the measurement system is greater than the repeatability standard deviation. See 8.5.2.5 for further information.

NOTE 2 If the precision changes significantly from one time to another, the assumption of identical distribution is not valid and this clause does not apply. Changes in precision can be detected by, for example, application of tests for homogeneity of variance or by inspection of residual plots. See 8.5.2.5 for further information.

8.5.2.2 Procedure

For the basic case, using simple linear regression, the procedure is as follows:

- Fit a preliminary model and inspect the fit and model residuals, checking any assumptions made (see 8.5.2.4 and 8.5.2.5 for further detail).
- If the assumptions apply, record the model parameters (usually slope and intercept) and their uncertainties and check the statistical significance of any trend found (see 8.5.2.6).

Detailed calculations for simple linear regression are given in B.3 and are generally available in all general-purpose statistical software packages and in most commercial spreadsheet applications.

NOTE Essentially all linear regression software, including commercial spreadsheet applications, carries out these calculations automatically. ISO 17034 requires that software, including any written by the RM producer for this purpose, be validated prior to use.

8.5.2.3 Model selection

For stability assessment where the underlying kinetic mechanism is unknown and changes are expected to be small, a linear approximation is usually a suitable model. The simple linear relationship is described in B.3.

In cases where a well-defined nonlinear mechanism is the reason for the instability, the corresponding degradation model is to be preferred over the (empirical) linear model. The mathematics is somewhat more complex for models other than the straight line, but the evaluation runs in the same fashion.

EXAMPLE An RM containing a radioactive isotope is an example of a property with a well-defined kinetic mechanism, in this case a radioactive decay, which can be predicted by a well-known but nonlinear model.

8.5.2.4 Fitting the model

The regression parameters can be computed using the procedures in B.3 or by suitable software. The calculations provide estimates b_0 and b_1, respectively, of the true intercept and slope β_0 and β_1, together with the corresponding standard errors $s(b_0)$ and $s(b_1)$, which can be used in subsequent statistical tests and in uncertainty evaluation.

8.5.2.5 Inspection and check of assumptions

The regression results should be inspected and the assumptions checked as indicated in B.3.3.

In reference material stability studies, it is particularly important to check for evidence of an inappropriate degradation model. In the case of a simple linear model, evidence of nonlinearity is therefore particularly important.

Where serious departure from the assumptions is found, data analysis should be discontinued and any anomalies should be resolved or alternative data treatment adopted.

Some common issues and courses of action are:

— Outlying data points should be inspected, and if appropriate, removed. Outlying data points adversely affect both model fits and tests for significance. If attributable to measurement system failures, outlying data points should be corrected or removed. However, outliers can also be due to incorrect choice of degradation model or to RM units that have degraded individually; due care should accordingly be taken to consider possible causes before removing any observations. Individual outlying observations among replicates on the same unit in a stability study are, if the material has already been shown to be homogeneous, likely to indicate measurement failure.

— If there is significant between-unit heterogeneity (compared with the available measurement precision), it is necessary to allow for this in statistical assessment; in particular, significance can be overstated if between-unit heterogeneity is not allowed for.

— Evidence of run-to-run variation in the measurement system should be addressed by adopting alternative data treatment (see 8.5.5).

— Evidence of curvature may be addressed using a different model or (since visible change implies instability) by concluding that the material is not sufficiently stable for the intended use.

8.5.2.6 Testing for significant instability

For the simple case above, the usual test for instability is a Student *t*-test for slope significantly different from zero as described in B.3.4.

If the trend is statistically significant, it is also useful to consider whether it is technically significant; that is, whether it is sufficient to require an increase of uncertainty in a certified value or to prevent certification. Trends should be considered technically significant if the predicted degradation over the period of validity of the material is important compared with the standard uncertainty of the property value in question.

If a technically significant trend is observed, the provisions of 8.6 apply.

8.5.3 Isochronous designs

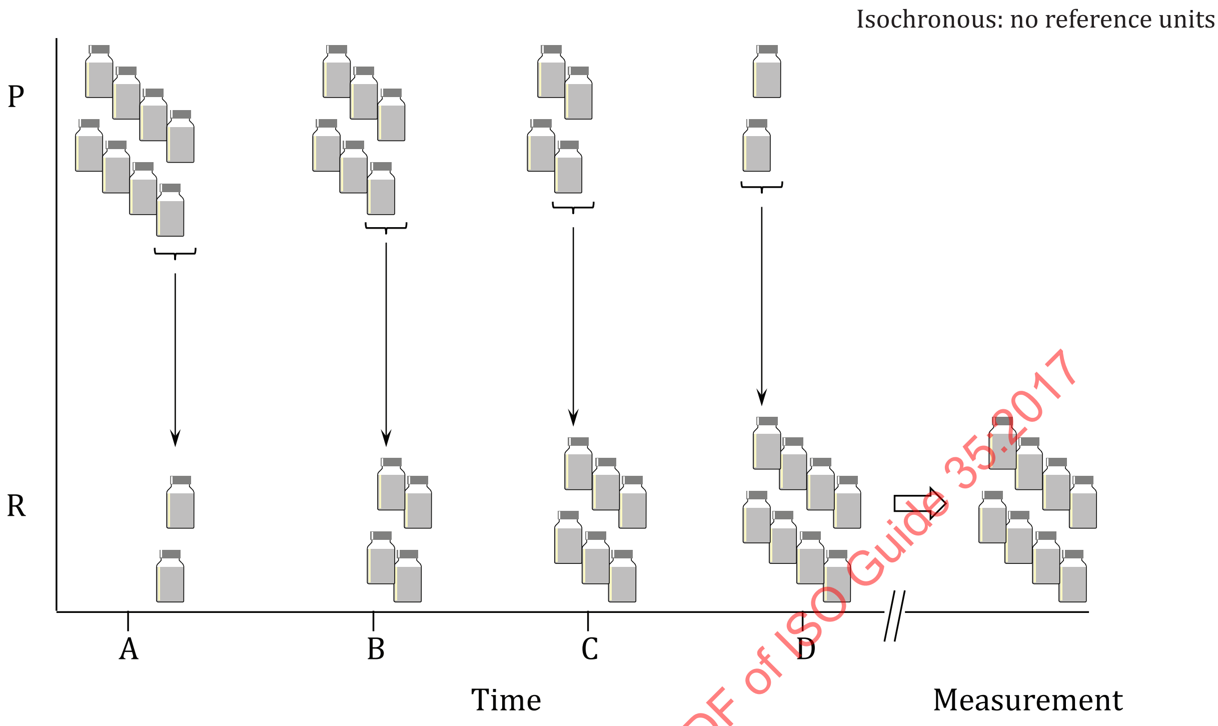
8.5.3.1 Simple isochronous study with one storage and one reference condition

In an isochronous design, a set of units of an RM is exposed to proposed storage conditions for a time, and then moved to reference conditions, after the planned exposure time for each unit. When all units have been exposed for the intended time, the complete set of units is measured. This sequence is illustrated schematically in Figure 6.

Usually, the planned storage condition is at ambient or near-ambient temperature (for example, 4 °C) and the reference conditions include a much lower temperature at which degradation is considered extremely unlikely.

Since the study involves only a single exposure condition, data analysis follows the procedure for the basic stability study at 8.5.2, using the time for which each unit is exposed to the planned storage conditions as the time axis for the analysis.

NOTE In an isochronous study, it is also possible to move material from reference conditions to planned exposure conditions to achieve the desired exposure time.



Key

P planned storage conditions

R reference conditions

NOTE The figure shows a schematic illustration of an isochronous study involving observations after exposure for increasing times A, B, C and D, equally spaced over the intended study period. A number of units (eight, here) are reserved for the study. At time A (which may be at a nominal zero exposure time), a set of units are moved to the reference conditions. An equal number are moved to reference conditions at each subsequent point in time. Finally, after all units have been exposed for the desired time, all units are stored under reference conditions and are then removed and measured simultaneously. Note that measurements need not be conducted under the reference conditions, provided that they are carried out within a short period of time.

Figure 6 — Illustration of an isochronous study

8.5.3.2 Advantages and disadvantages of isochronous stability studies

Isochronous studies, including the simple arrangement at 8.5.3.1, have the following advantages:

- All measurements are made in a short period of time, usually under repeatability conditions in a single run. This provides for the best available precision for the study and avoids the possibility that longer-term drift in the measurement system might be mistaken for instability.
- Completing all measurements in a single run reduces costs and simplifies the scheduling of laboratory resources.

The disadvantages include:

- Deferring measurement until the end of the study means that instability is not identified until after completion of the study.
- Simple interpretation of results in the basic design of 8.5.3.1 relies on minimal change under the reference conditions. If the material changes progressively under the reference conditions (for example owing to progressive freezing effects), this can either be mistaken for change at the planned storage conditions or can lead to an incorrect conclusion that the material is stable. Additional evidence can therefore be needed for reliable interpretation.

- The study design assumes that there is no adverse effect on transferring to, or from, the reference conditions. If the material changes on moving between conditions, the results can be highly variable or hard to interpret.
- Not all materials can be placed in conditions that are more effective at preventing degradation than the planned storage conditions. Some cannot be cooled to low temperatures; others can require storage at the lowest available temperature.
- Not all degradation mechanisms are affected by changed conditions, including temperature.

The basic design at 8.5.3.1 has the additional potential drawback of including only one candidate storage condition.

NOTE The disadvantages listed above can be reduced by extending the design as described in 8.5.3.3.

8.5.3.3 Extensions of the basic isochronous design

Some of the disadvantages listed in 8.5.3.2 can be addressed by extending the basic isochronous design. Extensions include:

a) Running two studies with different durations can provide earlier results, if desired.

b) Inclusion of multiple alternative exposure conditions, including ‘accelerating’ conditions. This can allow assessment of stability under a range of conditions and can additionally allow isochronous measurements for an accelerated stability study. In addition, including additional RM units retained under intermediate conditions (that is, between reference and planned storage conditions) provides some protection against misinterpretation of unexpected changes under reference conditions.

c) Measurement using randomized block designs. Where there are too many observations to complete in a single measurement run, the measurements can be carried out using a randomized block design. In this case, the effect of every exposure treatment is observed a smaller number of times in each of several measurement runs. This provides for a test of stability-related changes against the repeatability standard deviation, at the expense of a small loss in degrees of freedom. This is analogous to the use of randomized block designs in homogeneity studies (see 7.6.3).

Data analysis for the extended designs involved in b) and c) above is more complex than simple linear regression. For the treatment of a multiple-temperature isochronous study, the provisions of 8.5.4 apply. For randomized block designs, the procedures used are similar to those of 7.6.3 and generally require statistical software; however, a simple average of observations from each treatment across measurement runs permits simplified assessment as described above.

8.5.4 Accelerated stability studies with multiple exposure conditions

8.5.4.1 Description of accelerated studies

This type of study accelerates the occurrence of degradation effects from possible influential factors by submitting the objects under test to more extreme conditions than the planned storage condition, thus shortening the time within which an observable change occurs[21],[22] and/or increasing the observable degradation.

Factors that may accelerate degradation include, but are not limited to:

- environmental stress factors such as temperature, humidity, irradiation (usually light, especially UV), oxidising agents (e.g. air), storage container properties (contaminated contact surfaces, adhesive contact surface properties, leakage);
- sample-specific factors such as reactive matrix constituents, in particular for natural matrix RMs.

Since temperature plays a major part in at least the majority of known degradation processes, including chemical reactions, diffusion, evaporation, or adsorption, accelerated studies should normally test for the stress factor of temperature. The temperature range studied should include at least the range of temperatures the material might reasonably encounter under planned conditions of storage and transport.

Stress factors other than temperature, particularly humidity and irradiation, should be included in accelerated stability studies where there is no existing evidence that they can be neglected.

The following paragraphs are limited to the stress factor temperature, although the general approach can also be applied for any other continuous stress factor.

NOTE If no change is observed in an accelerated study then it is not possible to fit a reliable predictive model. It is therefore useful for subsequent mathematical modelling and prediction of change if the conditions included in an accelerated study can be chosen so that appreciable change is observed at least at the extremes.

For pharmaceutical products, temperature and humidity, and in special cases irradiation, can have an impact on the stability of the material. Recommendations of the International Conference on Harmonization (ICH) describe accelerated degradation studies in which the material under test is exposed to certain defined combinations of these factors[20].

EXAMPLE An example of a possible layout of an accelerated ageing study for a candidate material is given in Table 1. Depending on the expected behaviour and the rate of degradation, a layout might contain more or fewer stress points in time and temperature, and ranges may vary according to the kind of measurand under investigation.

Table 1 — Accelerated ageing example: Temperatures and exposure times

Temperature (°C)	4	20	40	70
Exposure time (months)	1, 3, 6, 12, (24)	1, 3, 6, 12, (24)	1, 3, 6	0,5, 1
NOTE Exposure times in parentheses show sampling times for the first post-certification monitoring (see 8.10)				

8.5.4.2 Mathematical models for assessing accelerated stability studies

This time dependence $g(t)$ of the temperature-driven degradation of an analyte under consideration is most commonly modelled using the basic assumption of proportionality between the rate of change of analyte with time (dx/dt) and the total amount of the analyte available in the sample. The proportionality coefficient is the temperature-dependent reaction rate $k_{\mathrm{eff}}(T)$. Thus, from Formula (9),

$$\frac{dx}{dt} = -k_{\mathrm{eff}}(T)x \tag{9}$$

with x being the value of the measurand under investigation, one obtains an exponential dependence on time. Although this assumption is not universally applicable, it describes a very large number of degradation processes. Simple linear modelling following this assumption should therefore refer to $\ln[x(t)]$ rather than the value $x(t)$ itself, and the regressions as mentioned below should be carried out in a log-linear space or should use nonlinear curve fitting appropriate to the model chosen.

At a given temperature, the effective pseudo first-order reaction rate $k_{\mathrm{eff}}(T)$ can arise from more than one concurrent degradation process, each with a different dependence on temperature. Temperature-driven degradation processes often follow an Arrhenius model, biological/microbial processes can follow an O'Neill model and diffusion-driven changes in a sample can follow Fick's law. If prior information on possible degradation processes is available, the dynamic reaction kinetics can serve as the basis for describing the dependence of the reaction rates on the level of experimental factors. Empirical models (without a physico-chemical rationale) are acceptable in cases when a larger number of individual processes overlap and can (at least partially) compensate.

An appropriately selected model is then regressed over the whole set of data (referring to all stress levels) from the stability study providing, for any stress level, an estimate for the model parameters. These estimates then will allow the RM producer to

— estimate a final expiry date for materials with detected instabilities (under storage conditions at the RM producer's site);

- estimate an uncertainty of stability for materials;
- estimate the maximum admissible re-testing interval for seemingly stable materials (post-certification monitoring);
- assess maximum admissible stress loads (temperature and time) during delivery of the material to the customer.

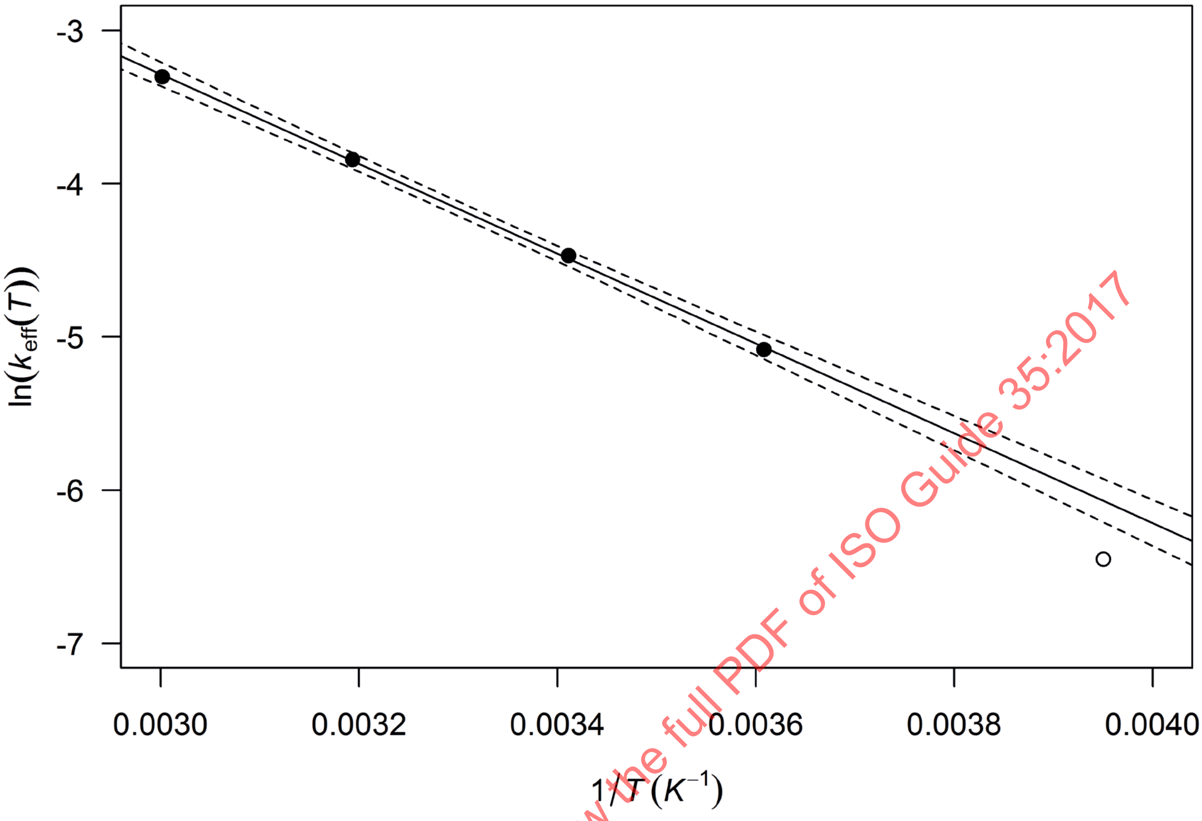
Subject to 8.4.3 and to appropriate study design (8.4.4), where no degradation is found at any stress level in an accelerated study in which the highest temperature is at least 20 °C above the planned storage temperature, the degradation rate under planned storage conditions may be assumed to be negligible, subject to confirmation by appropriate monitoring arrangements (see 8.10).

Extrapolation beyond the range of storage conditions tested (for example, predicting degradation rates at −20 °C from an experiment involving only temperatures above 0 °C) can be unreliable and is not recommended.

NOTE 1 Processes behind degradation/deterioration of a material can be complex and can require different models. In many cases, complex dependencies of the form $x(t)=F(T,t)$ can, for stability study purposes, be decomposed into a relationship of the form

$$x(t)=F(T,t)=f(T)g(t)$$

EXAMPLE Figure 7 shows an experimentally obtained dependency for total petroleum hydrocarbon (TPH) content in an environmental CRM.



NOTE The figure shows an Arrhenius plot for a CRM certified for total petroleum hydrocarbon (TPH) content in soil. The figure shows the dependence of the effective degradation rate $k_{\text{eff}}(T)$ (solid circles •) on the inverse temperature and the 95 % confidence interval for the line. Note that the final data point (empty circle ○) was excluded from the regression analysis for technical reasons; the regression line is nonetheless extended to show the predicted change at that temperature and better illustrate the confidence limits. The fitted relationship is good for the higher temperatures represented by the first four points.

Figure 7 — Results for an accelerated stability study

NOTE 2 More application examples are given, and comprehensively described, in References [23] and [24].

8.5.5 Additional sources of random variation in stability studies

Often, there can be more than one random effect; for example if multiple RM units are observed in duplicate at each point in time, there is (in principle) a between-unit term, as well as the residual error term. Specialised treatment, ideally using statistical software, is required to take full account of such an error structure. The most common approach is mixed-effects modelling using maximum likelihood methods, which allows for multiple random effects and calculates the variances for the random effects and the required standard errors for fixed effects (slope and intercept) accordingly[25],[26]. However, a simple treatment when the number of replicates per RM unit is identical is to average multiple replicates to give a single value per RM unit; this can then be treated as a statistical model with a single residual error term. The disadvantage is that this reduces the apparent number of degrees of freedom and can consequently increase the uncertainty associated with regression coefficients.

8.6 Action on finding of a significant trend in a stability study

If a technically significant trend (see 8.5.2.6) is observed, one of the following approaches should be adopted:

a) the property value for which the degradation was observed is not certified;

b) the period of validity of the certified value is decreased, based on the model prediction, to ensure that the change in value is not technically significant;

c) the expected extent of degradation over the intended period of validity is estimated, converted into a standard uncertainty and included, together with the uncertainty of the expected degradation, in the uncertainty of the assigned value;

d) the certified value and its uncertainty are given as a function of time, reflecting the estimated trend and its uncertainty;

e) a combination of two or more of b), c) and d).

NOTE 1 The approach in c) above is equivalent to including an uncorrected bias in a measurement uncertainty.

NOTE 2 Approach d) above requires a well-established degradation model.

NOTE 3 8.7.4 gives further details of the evaluation of uncertainties in the presence of a significant trend.

EXAMPLE 1 A CRM is certified for the concentration of ^{242}Pu in solution. This concentration decreases over time due to the well-established radioactive decay of the isotope. The certificate states that the value is valid for one exact day (the day on which the measurements were performed) and states the yearly loss of the isotope (0,000 22 % per year), with the half-life used and the uncertainty of the half-life, to allow calculation of an exact value on the date of use as well as an uncertainty for the value at that time.

EXAMPLE 2 An RM for quantitative microbiology is found to lose, on average, 0,05 log cfu per month of storage, with an uncertainty of this loss of 0,01 log cfu per month (as standard uncertainty), and the loss is considered important compared with the uncertainty of the property value. The initial property value is the value obtained at the time of characterization. To allow for loss on storage, certificates are issued with a value corrected for any lapse of time since characterization and the uncertainty of the correction is included in the claimed uncertainty for the property value (see 8.7.4).

8.7 Uncertainty evaluation from stability studies

8.7.1 General considerations for uncertainty evaluation from stability studies

Where valid technical reasons demonstrate that the potential change over the period of validity of the certificate is negligible compared with the certified uncertainty (e.g. less than $u_{\mathrm{CRM}}/3$), and this is supported by experience and observation, then the component of uncertainty due to long-term stability may be set to zero or omitted from the uncertainty in the certified value.

In other circumstances, where stability data analysis produces estimated rates of change, usually derived as, or from, coefficients in a fitted model, it becomes possible to predict potential future change based on the model. Such predictions can inform the choice of monitoring regime. In addition, it becomes possible to make an estimate of the uncertainty associated with predicted changes in certified values over time. Where little prior information is available about the behaviour of the material over extended periods of time (that is, larger than the stability study duration), and the RM producer chooses to employ comparatively infrequent monitoring (e.g. yearly or less), it is prudent to estimate these uncertainties and, for certified reference materials, to include them in the uncertainty associated with the certified value(s).

This subclause (8.7) provides basic guidance on the sources of uncertainty to be considered in these circumstances and on the general procedure for estimating uncertainty associated with possible change over time.

NOTE 1 Matrix degradation can affect measured values obtained by some measurement procedures even where the (true) value of the property remains stable.

NOTE 2 Judgement about stability based on valid technical reasons, supported by experience and observation, includes consideration of the size of possible changes in relation to the size of the uncertainty required for the intended use of the material.

8.7.2 Sources of uncertainty in predicted change over time

A stability study includes the following sources of measurement variation:

a) repeatability of measurement;

b) between-run variability of the measurement system;

c) between-unit heterogeneity (in batch characterization).

All the sources of measurement variation (a to c, above) contribute to uncertainty in prediction, with the exception that in isochronous designs run-to-run variability is only present where multiple measurement runs are necessary. Where these sources of variation are present and contribute to the variation in measurement results during the study, the standard errors in estimated model coefficients will, if properly fitted, take due account of these uncertainties.

Uncertainties associated with random variability during the study should be included in any estimate of the uncertainty associated with a predicted change over time.

Where a predicted change uses several model coefficients, the estimated uncertainty should take account of the complete covariance matrix for the coefficients used, as errors in model coefficients are often highly correlated.

In addition, systematic effects are present, including (but not limited to):

— uncertainties in model coefficients arising from measurement of time, or measurement of response;

— uncertainty arising from the choice of model, for example an assumption of linear degradation rather than exponential change or (more severely) instead of an autocatalytic model.

Contributions associated with systematic effects on the measurement of time and response are usually much smaller than those arising from random variation. It is not common practice to include these contributions in estimating prediction uncertainties used solely for uncertainty evaluation, though it is prudent to do so when estimating a correction to be included on a certificate or in instructions for use.

The model used for extrapolation from the conditions of the accelerated study to the effects in real time under defined storage conditions should be valid, and the uncertainty of this model should be considered. Uncertainties associated with choice of underlying model are not usually included in estimating prediction uncertainties. Instead, alternative candidate models may be fitted and their predictions used to inform the selection of the monitoring regime.

NOTE 8.5.5 discusses the treatment of multiple sources of variation in stability studies.

8.7.3 Estimation of stability uncertainties in the absence of significant trends

When there is technical justification for stability (see 8.7.1), it may be assumed that the material is stable and the uncertainty associated with stability may be set to zero along with adopting a monitoring regime that can detect unexpected change promptly, ideally before it can adversely affect the use of the material. The choice of initial monitoring point and subsequent intervals are discussed in 8.10.

Where there is no technical justification for stability that is supported by experience, and where the RM producer elects to use longer monitoring intervals, an uncertainty associated with possible instability should be estimated and included in the uncertainty associated with any certified value. The uncertainty associated with possible instability should then be based on the uncertainty associated with prediction of the change in value at a time equal to the first monitoring point plus the time to expiry of any certificate issued up to that time. The prediction should use an appropriate model fitted to the available stability data and the uncertainty should take due account of all relevant sources of random variation (see 8.7.2).

For a simple linear model applied to a classical stability study over several points in time, the uncertainty u_{lts} associated with the predicted change is given by

$$u_{\text{lts}} = s(b_1)(t_{m1} + t_{\text{cert}}) \tag{10}$$

where $s(b_1)$ is the standard error for the estimated slope, calculated as in B.3, t_{m1} is the time interval between value assignment and the initial stability monitoring point and t_{cert} is the period of validity of a certificate issued during that time.

Some RM producers do not provide an expiry date specific to each certificate issued, preferring to set an expiry date based on a planned lifetime for the material. In such cases, the term (t_{m1} + t_{cert}) in Formula (10) may be set to the planned lifetime (that is, t_{cert}) or, at the RM producer's discretion, to the time to the second planned monitoring point, relying on immediately informing users of any change discovered after the initial monitoring point.

NOTE The justification for Formula (10) is given in Reference [18].

8.7.4 Evaluation of stability uncertainties in the case of a known significant trend

Where there is a known statistically significant or technically significant trend the RM producer may, following 8.6, provide a time-dependent certified value or may provide a time independent certified value and increase the prediction uncertainty. This clause provides further detail on the evaluation of uncertainty in these two cases.

When a time-dependent value is provided, the function for the certified value should reflect the best estimate of the trend and the function for the uncertainty should reflect the uncertainty of the trend (or, if appropriate, a correction), taking into account the sources of uncertainty listed in 8.7.2.

When a certified value is given independently of time, the RM producer should increase the prediction uncertainty to allow for expected change. A period of validity (usually (t_{m1} + t_{cert}) as in 8.7.3) is chosen and the extent of degradation over that time is estimated. This is converted into a standard uncertainty (e.g. using a rectangular distribution if one observes a linear trend) and this uncertainty is combined (using uncertainty propagation rules) with the uncertainty of the predicted change.

NOTE The resulting uncertainty for a time-independent certified value can be applied only to one side of the interval, as degradation tends to be in one direction only, thus resulting in asymmetric uncertainties.

8.8 Estimation of storage lifetime ("shelf life") from a stability study

Where it is possible to set an acceptable level of change due to lack of stability for a certified value, it is possible to estimate a storage lifetime within which the value is expected to remain acceptable for use. The principles are described in B.4.

NOTE An acceptable level of change can be set from, for example, a specification limit

8.9 Instructions for use related to management of stability

The instructions for use of the material should include detailed information on how the material should be stored by the user to ensure that the material does not deteriorate beyond the stated uncertainty during the period of validity. This information should include recommended storage conditions prior to and (explicitly, if the material may be used more than once) after opening. Any necessary restrictions on time of use after opening a unit should be included. The instructions for use may include instructions for verification of the integrity of a reference material prior to first use; such verification might, for example, be based on package seal inspection or temperature logging equipment included in the packaging.

Where repeated subsampling is permitted, the instructions for use should include any precautions for prevention of contamination and for storage of opened units of the RM that are necessary to ensure that the remaining material remains fit for use and, for CRMs, that the stated uncertainty is not compromised. If property values can be affected by repeated subsampling, for example, by evaporation or by repeated refreezing, this should be noted on the certificate.

NOTE Instructing the end user to store reference materials away from light, away from sources of heat and in dry conditions is good practice. Such instructions also minimize the range of conditions that require experimental study to verify required storage conditions.

8.10 Stability monitoring

8.10.1 Requirements for monitoring

Monitoring should be planned for during the lifetime of the RM. As noted in 8.1, the behaviour of a given RM over its lifetime is difficult to predict reliably from typical stability studies. Because the behaviour is hard to predict, it is usually necessary to monitor the stability of the material. Monitoring of a material following release is accordingly an important part of the overall management of stability for reference materials with long usable life.

If experience of stability from previous RM production batches is to be used to inform stability claims for future batches, including shelf life, rates of degradation or uncertainty due to long-term storage, the RMP should have evidence to support claims of stability. In these cases, the monitoring tests should include a measurement made at the expiry date for previous batches.

Monitoring is, however, not always essential. Some materials are certified for use with very short lifetime and early expiry dates, with replacement materials in near continuous production. Such materials expire before any reasonable monitoring point. The RM producer should therefore assess the need for post-release monitoring. The assessment should consider:

- the duration and results of stability studies;
- experience of the stability of related materials (or previous batches of essentially identical material);
- the expected sales lifetime or availability lifetime of the material;
- whether allowance for change is to be included in any statement of uncertainty for the properties of interest;
- the strength of other evidence, for example from literature studies, for the assumption that the property value(s) will remain stable.

Monitoring is not necessary where the expected lifetime is short compared with known degradation rates for the same or closely similar materials. In most other cases, some monitoring is normally considered necessary and should be undertaken at least once over the lifetime of the material to confirm stability.

8.10.2 Choice of initial monitoring point and monitoring interval(s)

8.10.2.1 Relevance of prior information

Where monitoring is considered necessary, the RM producer should set an appropriate initial monitoring point and intervals for further monitoring. Both choices depend heavily on the availability (or otherwise) of information acquired over extended periods on closely similar materials.

8.10.2.2 Monitoring plans where extensive prior information is available

Where there is sound, relevant information from stability studies and/or monitoring of closely similar materials, covering a period similar to (or longer than) the expected lifetime of the material in question, the RM producer may set an initial monitoring point and intervals that are

— similar to those used successfully on closely related materials, or

— based on the observed change over time for the previous materials.

In the latter case, intervals should be set so that reasonably expected change between monitoring points, based on prior information, is not more than one third of the uncertainty associated with certified values.

In assessing the possible change, the uncertainty of any experimental assessment should be taken into account, for example, by selecting shorter intervals when the uncertainty is large. Relevant prior information from published literature, experience or other sources may also be taken into account in assessing the possible change over time and for selecting a monitoring interval.

8.10.2.3 Monitoring plans where extensive prior information is not available

8.10.2.3.1 General considerations for monitoring where prior information is not available

Where there is insufficient information from stability studies and/or monitoring of closely similar materials, and in particular, where no similar material has been available for longer than the duration of stability studies carried out on the material(s) in question, the RM producer should set initial monitoring points based on the stability study results for the material in question, and in addition plan for comparatively frequent monitoring at least for the first three monitoring points.

There are two basic strategies for choosing initial monitoring points in the absence of extended data on closely similar materials:

a) predict possible change and set the initial monitoring point prior to any change that adversely affects end use;

b) use a simple multiple of the stability study duration.

8.10.2.3.2 and 8.10.2.3.3 provide further details of these strategies.

In both cases, a) and b), intervals for subsequent monitoring points (that is, after the first three points) should be set following review of the results of measurement at the first three monitoring points.

NOTE Strategy b) is not appropriate if the long-term stability study time, t_{lts}, is long compared with the expected lifetime of the material.

8.10.2.3.2 Use of a predicted change to set the initial monitoring point

To use a predicted change to set the initial monitoring point and interval [strategy a) in 8.10.2.3.1], a specified tolerance for the certified value(s) is required. The procedure is straightforward in principle. First, a two-sided confidence interval for the change in certified value is constructed for a series of times following value assignment. Second, the earliest point at which one of these limits intersects the limits of the specified tolerance is determined, either graphically or numerically. This point, or a convenient earlier time, is taken as the first monitoring point. Details are given in B.4.

NOTE The specified tolerance for a reference material can be based on considerations for intended use. For a CRM, such an interval is usually based on the expanded uncertainty. For example, choosing a tolerance of one third of the expanded uncertainty gives a low risk of the certified value moving outside of the expanded uncertainty prior to the first monitoring point.

8.10.2.3.3 Use of a simple multiple of the stability study duration

This strategy [strategy b) in 8.10.2.3.1] is simple but can result in shorter intervals than strategy a). It is based on multiples of the long-term stability study duration t_{lts}. The multiple used should be chosen to limit the risk of a change that might affect the end use before the first monitoring point.

EXAMPLE An example of an application of strategy b) is given below.

— Set the first monitoring point at the later of the value assignment date plus t_{lts} and the date of the end of the long-term stability study plus t_{lts}.

— Set two subsequent monitoring points at intervals of $2t_{\text{lts}}$ from the first.

8.10.3 Experimental approaches and evaluation for stability monitoring

8.10.3.1 Classical monitoring design

Monitoring often takes place using the classical design. This involves measurement of RM units in normal storage at planned points in time. The evaluation of the results involves a comparison of each (mean) monitoring result with the certified value and, over time, a check for any significant trend in the observed value.

The advantage of the classical design is simplicity; the principal disadvantage is that the results can be adversely affected by long-term variations in the measurement process.

8.10.3.2 Evaluation of stability monitoring results

The basic evaluation of a single stability monitoring experiment applied to a CRM relies on comparison of the new measured value with the certified value. The approach requires the standard uncertainties u_{mon} and u_{CRM} associated with x_{mon} and x_{CRM}, respectively, and an appropriate coverage factor k at a level of confidence of approximately 95 %. Using this method, if the condition

$$\left|x_{\text{CRM}} - x_{\text{mon}}\right| \leq k\sqrt{u_{\text{CRM}}^2 + u_{\text{mon}}^2} \qquad (11)$$

is not met, then it should be concluded that there is evidence of instability.

Where previous monitoring results on the same value are available in addition to the certified value, a check for a trend in the values should be performed. A check for a trend in the values over two or more monitoring points (in addition to the certified value) may be performed using simple linear regression. Weights may be applied if the uncertainties at different points differ appreciably. Where the gradient is significant at the 95 % level of confidence, it should be concluded that there is evidence of a trend in the values.

Where the criterion in Formula (11) is not met or where there is evidence of a trend, this indicates significant degradation of the material and action should be taken. Possible modes of action can include:

— performing confirmatory studies (with or without temporary suspension of RM distribution);

— halting distribution and discarding the material;

— re-certification of the material.

Following evaluation, new monitoring data may be added to the collected stability data for the material in order to revise future monitoring points or amend the estimated period of validity of the certified value.

NOTE Meeting the criterion in Formula (11) cannot, in isolation, be taken as positive proof of stability, especially future stability, or that no change at all has occurred. It can only be concluded that the change, if any, was not large enough to detect with the available uncertainty.

8.10.3.3 Isochronous stability monitoring

An alternative to classical monitoring is to use a variant of the isochronous experiment design to carry out a type of semi-continuous stability study. This involves storing some units of the material under reference conditions at which the material is expected to be extremely stable, typically at very low temperatures. At each monitoring point, units kept under normal storage conditions are measured together with units from the reference conditions under (ideally) repeatability conditions of measurement. Data processing should include one or more of the following:

— a test for significant change compared with the units stored under reference conditions;

— inspection and analysis for trends in the difference or ratio in value between samples in normal storage and samples held under reference conditions;

— modelling that allows for run-to-run variation in the measuring process.

NOTE Reference [18] suggests an alternative scheme in which fresh RM units are placed at the reference temperature at each monitoring point. This avoids the necessity of a large stock of RM units held at the reference temperature for possible future monitoring requirements, at the expense of more intricate data analysis.

9 Characterization of the material

9.1 Preamble

The guidance in this clause is intended mainly for the measurements performed to assign the certified property values of a material (y_{char}). Studies for the determination of non-certified property values (however named), may also follow the principles outlined in this clause, but will in general require less rigour, especially with respect to evaluation of measurement uncertainties and establishment of metrological traceability (see 9.11).

It is important to note that a certified property value should be a good estimate of the true value and not just the average of a population. The certified value may be the same for many individual units (batch processing), or an individual value may be assigned to each unit in cases where a number of single artefacts are being produced.

For certified values, the associated uncertainty of characterization (u_{char}) should be determined. ISO 17034 requires an RM producer to provide *evidence of the metrological traceability of the certified value to a stated reference*. This means that whatever the approach chosen, the metrological traceability of the certified values should be clearly defined. Traceability can only be achieved if the values that are combined have been shown to provide valid estimates of the value of the measurand (as defined) within the claimed uncertainty and the results are traceable to the same metrological reference. Ideally, the International System of Units (SI) is the preferred metrological reference, but other references can be used. Metrological traceability also applies to operationally defined quantities; it remains essential to ensure traceability to defined metrological references by proper calibration.

NOTE 9.2 gives guidance on establishing metrological traceability for reference material characterization.

Characterization can be achieved by using one or several methods in one or several laboratories[27]. ISO 17034 lists several basic approaches to characterization:

— using a single reference measurement procedure (as defined in ISO/IEC Guide 99) in a single laboratory;

— characterization of a non-operationally defined measurand using two or more methods of demonstrable accuracy in one or more competent laboratories;

— characterization of an operationally-defined measurand using a network of competent laboratories;

— value transfer from a reference material to a closely matched candidate reference material performed using a single measurement procedure performed by one laboratory;

- characterization based on mass or volume of ingredients used in the preparation of the reference material.

This clause provides guidance on these basic principles, as well as on the conduct of collaborative studies, characterization of purity and characterization by direct comparison with closely matched CRMs. While the current state-of-the-art is not sufficiently evolved to give detailed guidelines for the characterization of qualitative (nominal) properties (e.g. identity of the substance), some general principles are also listed in this clause.

9.2 Establishing metrological traceability

9.2.1 Principle

Metrological traceability is a characteristic of a measurement result. In practice, the traceability of a measurement result of a property value consists of two parts, namely the clearly defined identity of the measurand and the traceability of the property values of this measurand to the stated references. Establishing traceability therefore includes both the proof of identity of the property measured and the comparison of the results to an appropriate stated reference. The comparison is established by ensuring that measurement procedures are properly validated, that measuring equipment is appropriately calibrated, and that any conditions of measurement (such as test material preparation, environmental conditions, etc.) are under sufficient control to provide a reliable result. An RM producer can ensure this in a number of ways, including validation of procedures and calibration of equipment under their control, or verification of traceability through the use of materials of known value. The following clauses give further guidance on these principles.

NOTE Further guidance on metrological traceability for chemical measurements is given in References [28],[79] and [80].

9.2.2 Metrological references

The traceability of measurement results is usually ensured through proper calibration of all relevant input quantities against appropriate measurement standards and/or certified reference materials. Quantity values can be traceable to

- a generally accepted system of units [e.g. the SI];
- measurement standards, including CRMs.

In most cases, laboratories will use measurement standards that carry values traceable to a higher reference (e.g. SI). This should be attempted wherever possible. Values obtained by calibration with such a standard are traceable to this higher reference via the standard, if all other input factors have been duly calibrated.

NOTE 1 Typical items to be calibrated include balances, thermometers, torque-wrenches, volumetric instruments, vernier calipers and stopwatches.

NOTE 2 In many cases, measurement standards, including CRMs, will be used as calibration standards for the measurement procedures. Examples are working standards (traceable to the primary standard) or conventional scales like pH, for which the agreed primary realization is the Harned cell and for which routine calibration uses buffer solutions. Values obtained by calibration with this standard are traceable to this higher reference via the values of the standard, if all other input factors have been properly calibrated.

NOTE 3 If the certified value of a CRM used for calibration is itself traceable to a higher reference (e.g. the SI), then the new CRM will be traceable to this higher reference via calibration with the CRM, if all other input factors have been properly calibrated.

NOTE 4 The relevance of input quantities is usually evaluated against the combined standard uncertainty of the measurement result. A common rule of thumb is to consider the uncertainty contribution of one input quantity relevant if it is larger than a third of the combined standard uncertainty.

9.2.3 Types of measurands

A measured property can be

- defined without reference to a particular procedure for measurement. This is the case for basic physical properties (length, mass) and concentrations of clearly defined substances, which can be directly linked to the amount of substance (mole). In this case, the measurand is meaningful without reference to a particular measurement procedure.

- operationally defined. In this case, the measurand is defined by reference to a documented and widely accepted measurement procedure and only results obtained by the same procedure can be compared.

Whether or not the measurand is operationally defined, the establishment of traceability requires the same activity; every quantity that materially affects the measurement result should be subject to calibration or should be kept under suitable control, usually by use of calibrated instruments[81].

EXAMPLE 1 The concentration of Cd in a sample is to be certified. The measurement procedures chosen are two validated procedures based on acid digestion followed by ICPMS and neutron activation analysis. The measurements are calibrated using certified solutions of cadmium. The results from two very different principles of measurement agree within their respective uncertainties. Together with the validation data and evidence of calibration, this provides confidence that bias for either procedure is small compared with the uncertainty showing that the measurand is not operationally defined and that the certified value is traceable to the stated reference.

EXAMPLE 2 The mass fraction of crude fibre as defined by ISO 6865[29] is determined by an interlaboratory study in which all participants apply ISO 6865. All measurement conditions (temperatures, volumes, mass, etc.) were properly calibrated. The measurand is operationally defined (ISO 6865) and results are traceable to the SI.

EXAMPLE 3 The mass fraction of crude fibre is determined by near-infrared spectrometry (NIR). The instrument is calibrated using measurement results obtained following ISO 6865. The measurand is operationally defined (mass fraction of crude fibre as determined by NIR) and the quantity values are traceable to the results of ISO 6865.

NOTE The results from the NIR procedure will usually have larger uncertainty than the defining procedure because the uncertainty must include the uncertainties of the calibration values as well as additional uncertainties arising from use of the NIR procedure, including allowance for possible procedure bias on the particular material.

9.2.4 Effect of sample preparation or pre-treatment

For many matrix reference materials, the situation is complex. Although the instrumental determination of the property value can be made traceable to appropriate units by the calibration of the measurement equipment used, pre-treatment steps such as extraction, pre-conditioning or transformation of the sample from one physical or chemical state to another cannot easily be calibrated. Such treatments can only be compared with a reference procedure (when available), or among themselves. This makes the clear definition of the measurand somewhat complicated. Generally, three possibilities exist:

a) For some treatments, reference measurement procedures have been defined and may be used in characterization studies to provide a certified value defined by reference to the reference measurement procedure. This gives an operationally defined quantity.

b) A second possibility is the use of two or more independent procedures to assess the procedure bias. If the results from independent procedures agree within their respective uncertainties, the RM producer may conclude that the values obtained are not significantly influenced by the individual procedures and hence the measurand is not operationally defined.

c) In other cases, only a comparison among different laboratories using the same procedure is possible. In this case, it is impossible to demonstrate absence of method bias; therefore, the result is an operationally defined measurand.

EXAMPLE The mass fraction of Cd in soil was determined in several laboratories that all used aqua regia extraction and subsequent quantification by ICPMS. The measurand is operationally defined as “obtained by aqua regia extraction and subsequent quantification by ICPMS”.

The definition of the property on the documentation provided to the users should reflect the characterization approach chosen.

9.2.5 Verification of traceability

In many cases, it is difficult to demonstrate the proper calibration of each and every piece of equipment. Such situations can arise because of unknown influence factors, but arise more frequently in characterization studies involving multiple laboratories, where obtaining calibration certificates for each and every instrument used is impractical. In these cases, the adequacy of the measures taken to ensure proper calibration of equipment and the traceability of results should be verified by, for example, specially designed and prepared control samples (such as a sample otherwise used for calibration) and CRMs. Agreement of results on quality control samples can be used as demonstration of sufficient calibration of all relevant input factors.

NOTE Evidence of conformance with ISO/IEC 17025, including evidence from third party assessment, can be taken as additional evidence of traceability of the results reported by different measurement laboratories.

9.3 Characterization using a single reference measurement procedure (as defined in ISO/IEC Guide 99) in a single laboratory

9.3.1 Characterization by a reference measurement procedure without direct comparison with a CRM of the same kind

9.3.1.1 Concept

In this approach, a value is assigned by one laboratory using only one measurement procedure without direct comparison of a closely matched CRM. This limitation on the number of procedures and laboratories greatly limits the possibility to detect unexpected effects. Therefore, this approach requires the availability of a measurement procedure that is sufficiently well understood that unknown effects can be ruled out.

NOTE “CRM of the same kind” refers to a CRM which matches the CRM to be characterized in all characteristics that might have an influence on the measurement result (matrix, measured property, quantity value of the measured property, etc.).

The results and their uncertainty are then used to determine the assigned value.

9.3.1.2 Measurement procedure requirements

Any measurement procedure used for this approach should fulfil the following requirements:

- it is completely understood, meaning that all steps have a sound theoretical foundation so that systematic error is negligible relative to the intended use;
- it is completely described by a measurement equation containing all relevant influence factors linking the measurand to the properties actually measured, all of which can be expressed in SI units;
- the measurement equation does not contain empirically determined factors that have a major influence on the measurement result (e.g. “recovery rates”);
- there is no relevant influence of the measured quantity on any of the influence factors contained in the equation;
- the constants contained in the equation are known with a low uncertainty, which can be expressed in SI units;

- a realistic uncertainty budget can be written down in terms of SI units based on the individual quantification of the influence factors contained in the equation;
- the measurement uncertainty of the results obtained by the measurement procedure is sufficiently small for the intended use of the RM.

Establishment of the above requirements should be demonstrated by, for example, third party assessment, appropriate validation studies and measurement uncertainty evaluation in accordance with ISO/IEC 17025, verification of performance by comparison with other laboratories, proficiency tests, and so on.

The assigned value is the result obtained by the reference measurement procedure. The standard uncertainty of the result is expressed as u_{char}.

In addition to measurements with the reference measurement procedure, it is highly recommended to perform confirmation measurements with an independent measurement procedure to confirm the absence of gross errors. While confirmation by an independent measurement procedure is not strictly necessary, it is nevertheless highly advisable to provide additional confidence in the results, even if the confirmatory measurement results have a higher uncertainty than those from the reference measurement procedure. The confirmatory procedure can also be used to demonstrate the applicability of the material to measurement procedures other than the reference measurement procedure used for characterization.

Results of potential confirmation measurements do not need to be combined with the results of the reference measurement procedure, as their uncertainty is generally much higher. Instead, the results from the different procedures are tested to determine whether the results of the independent procedure agree with those from the reference measurement procedure. If this is the case, there is no evidence of method bias. If this is not the case, the cause (either a bias in the confirmation procedure or an unexpected effect in the reference measurement procedure) should be identified and the result corrected, if necessary.

9.3.2 Characterization by value transfer from a reference material to a closely matched candidate reference material using a single measurement procedure performed by one laboratory

9.3.2.1 Principle

In this approach, values are assigned to a “secondary CRM” by directly comparing results on the candidate CRM with those on an already characterized and closely matched CRM (the “primary CRM”). Examples for such materials include trace element solutions measured against certified solutions, materials measured against Pharmacopoeia standards or absorbance standards measured against certified absorbance standards.

NOTE 1 Each measurement on a candidate CRM that requires calibration in fact compares it with another CRM (the calibrator). This clause deals entirely with the case where the two CRMs are so closely matched that a direct comparison in one laboratory using one method can be sufficient to assign a certified value. Using a CRM of a different kind (e.g. a pure solution for calibration of measurements on a matrix material) falls into one of the other characterization approaches.

For this approach, the secondary CRM should be sufficiently closely matched to make material-specific bias negligible, within the claimed uncertainty, from the primary CRM in all characteristics which have a significant influence on the measurement result. In this respect, the following aspects should be considered.

a) The primary and secondary CRMs consist of the same matrix. Small differences to allow for the establishment of calibration curves are acceptable, as long as the main characteristics of the material remain unchanged. The primary and secondary CRMs present the same analytical challenges for the method used.

EXAMPLE 1 It is possible to characterize a solution of Cd in HNO_3 against a certified solution of Cd in HNO_3, as the matrix (diluted HNO_3) of both CRMs is the same. It is not possible to characterize Cd in granite as a secondary CRM against a solution of Cd in HNO_3 as the matrix differs. For many methods, the digestion step also adds an additional analytical challenge.

EXAMPLE 2 If chromatograms of secondary CRMs show co-elution with the analyte of interest, it is impossible to characterize them as secondary CRMs. If the primary and secondary CRMs differ in complexity (e.g. multi-component secondary CRMs characterized against a series of unmixed single-component primary CRMs), the RM producer should demonstrate that this added complexity does not influence the result.

b) If the measurand is not operationally defined, the matrix is of a kind that, for the measurement in question, the measurement procedure can be regarded as completely understood.

EXAMPLE 3 The chromatographic determination of a solution of benzo[a]pyrene is sufficiently understood, as no co-elutions or matrix effects occur. Determination of the benzo[a]pyrene in soil (or, in a soil extract) is not fully understood, as a multitude of factors can influence extraction efficiency and many co-elutions can occur.

c) The difference in the quantity level of the measured property does not result in a significant bias between the measurement results of the primary and secondary CRM.

NOTE 2 For chemical measurements, these conditions practically restrict the production of secondary CRMs to pure substances, solutions/dilutions of pure substances or operationally defined properties.

The measurement procedure used for characterization should fulfil all criteria for traceability listed in ISO 17034 and address the measurand for which the primary CRM is characterized.

The RMP should demonstrate the validity of the value and uncertainty transfer from the primary to the secondary CRM.

9.3.2.2 Assigned value and u_{char}

The assigned value is calculated by direct comparison between the results obtained on the primary and secondary CRMs. Valid methods include bracketing, multi-point calibration curves with the primary CRM, one point calibration with a primary CRM of closely matched certified value and adding the measured difference to the certified value.

u_{char} consists of a combination of the uncertainty of the certified value of the primary CRM, the uncertainty of calibration according to the chosen calibration model (which includes contribution due to the selectivity of the technique), and the effect of repeatability on the results of the secondary CRM. The calculated uncertainty should take account of the particular statistical treatment used to obtain the assigned value.

9.3.2.3 Traceability

The certified values of the secondary CRM are traceable, via the primary CRM, to the same reference as the values of the primary CRM.

EXAMPLE A solution of Cd in HNO_3 (secondary CRM) is characterized by measurement against a certified solution of Cd in HNO_3 (primary CRM). The certified values of the primary CRM are traceable to SI units. Therefore, the certified value of the secondary CRM is traceable to SI units as well.

9.3.3 Selection of RM units for single-laboratory characterization

The RM producer should use a measurement scheme (number of RM units, number of replicate results, etc.) that is capable of achieving the intended uncertainty for each certified value.

9.3.4 Formulation methods

This approach is usually applied for the production of calibration solutions from pure substances and also for gas mixtures, the production of which is described in a separate standard[30]. The approach is sometimes also used in the production of matrix materials.

The value of the measurand and its uncertainties, in all materials to be mixed, has to be known in order to calculate a certified value and uncertainty. In many cases, this is equivalent to determining the purity of the material of interest (see 9.6) and confirmation of the absence of the material of interest in the material to which it is added (for example, a solvent or 'blank' matrix material).

It is important to guard against change in content between acquisition and mixing; for example, water loss or uptake should be excluded, where appropriate.

If gravimetric mixtures of several materials, all of which contain the measurand in question, are to be prepared, each of the materials should be characterized using one of the approaches described in this clause.

Volumetric production follows similar principles in the calculation of the assigned value and uncertainty but entails an additional need to pay close attention to non-additive volumes in mixing liquids (for example, ethanol/water mixture volumes are not the simple sum of the water and ethanol mixed) and other factors affecting measured volume, particularly temperature.

For the case of purely gravimetric production to certify a mass fraction, the assigned value y_{char} is calculated from the masses m_i of the individual components and the mass fractions w_i of each material as in Formula (12):

$$y_{\text{char}} = \frac{\sum w_i \cdot m_i}{\sum m_i} \tag{12}$$

For the same procedure, the uncertainty u_{char} comprises all the uncertainties from the weighing steps as well as the uncertainties of the individual mass fractions. For the most common case of mixing two substances, this may be calculated from:

$$u_{\text{char}} = \sqrt{\left[\frac{m_1}{m_1+m_2}\right]^2 u_{w_1}^2 + \left[\frac{m_2\left(w_1-w_2\right)}{\left(m_1+m_2\right)^2}\right]^2 u_{m_1}^2 + \left[\frac{m_2}{m_1+m_2}\right]^2 u_{w_2}^2 + \left[\frac{m_1\left(w_2-w_1\right)}{\left(m_1+m_2\right)^2}\right]^2 u_{m_2}^2} \tag{13}$$

NOTE 1 Formula (13) does not include the covariance terms for correlated errors in weighing or the determination of w_i. Positive correlation effects can result in an increase in the combined uncertainty.

Although the result from the production is in principle sufficient for value assignment, it is highly advisable to check the result of the gravimetric production by measurement to detect any mistakes in the processing steps.

NOTE 2 A comprehensive discussion of the use of gravimetry in gas analysis is found in the literature (see, for example, References [31] and [32]).

9.4 Characterization of a non-operationally defined measurand using two or more methods of demonstrable accuracy in one or more competent laboratories

9.4.1 Concept

9.4.1.1 For many measurands, no reference measurement procedures are available that provide accurate results at the appropriate level of uncertainty. In these cases, it is necessary to find other means of improving the reliability of the assigned value. The approach described in this clause uses a number of data sets, obtained using different measurement procedures and/or in different laboratories to

a) demonstrate absence of significant bias in measurement procedures by showing that independent procedures yield the same results;

b) demonstrate the absence of significant laboratory bias for each laboratory by agreement among results;

c) improve the reliability of the assigned value by averaging results, thus reducing the effect of repeatability and randomizing and reducing the effect of between-laboratory or between-method variation.

9.4.1.2 The concept of the determination of the method-independent property values of an RM based on agreement among different measurement procedures, potentially performed in different laboratories, is based on at least two assumptions:

a) There exists a population of procedures and/or laboratories that is capable of determining the characteristics of the RM and providing results with acceptable accuracy.

b) For most data evaluation approaches, it is assumed that the differences between individual results, both within and between measurement procedures/laboratories, are random in nature regardless of the causes (for example, variation in measurement procedures, personnel or equipment).

9.4.1.3 For this approach to be valid, all results of all measurement procedures and/or laboratories involved should determine the same measurand and the results should be traceable (see 9.2.2) to the same system of units. This requires careful selection of calibration standards and careful investigation of the measurement procedures used. 9.4.2 describes selection of calibration standards.

NOTE 1 Even at the "state-of-the-art" level, differences in performance characteristics of measurement procedures as well as differences in the magnitude of uncertainty can exist between laboratories.

NOTE 2 There can be different objectives of interlaboratory comparisons, among them method validation, proficiency testing and characterization of reference materials. The goal of the study has important implications for the setup and evaluation of the various studies. It is therefore important to keep the goal of characterization in mind and not to mix it with other purposes, even if it is logistically combined with, for example, a proficiency testing exercise (see A.3).

9.4.1.4 Inter-laboratory and multiple-method characterization rely in part on averaging across different sources of bias, to achieve a reduction in uncertainty. Effective averaging relies on representative sampling for different effects. This has important implications for the choice of participants and measurement procedures:

— Where possible, measurement procedures should be selected to give a good representation of different principles of measurement.

— The choice of participants should be representative of competent laboratories.

The choice of measurement procedures is discussed in A.1.1. The representative selection of participants is discussed in A.1.3.

9.4.2 Study design

At least two substantially different measurement principles should be included in a multiple-method study. For interlaboratory studies using many participants with free choice of measurement procedures, a good representation of measurement procedures suitable for the determination of the particular characteristic should be sought.

Consideration should be given to the choice of the calibration standard, i.e. whether each participant should use a standard of its choice or whether a common calibrant is provided to all participants. The purity of the calibrants used should be given due consideration.

Laboratories should be selected based on demonstrated competence. Therefore, participating laboratories should provide evidence of competence for the measurand in question independent of the measurements on the candidate CRM, ideally before commencement of the study. It is thus impossible to use data on the candidate CRM from the same study as demonstration of competence and for value assignment of a CRM (e.g. using the consensus value of results of a proficiency test study for value assignment of a CRM). (See also A.3.)

The RM producer should set a documented minimum number of technically valid results for which value assignment will be considered. The number of data sets should be large enough to provide a fit-for-purpose uncertainty in the estimated value after allowing for the possibility of failure to report, exclusion of results for technical reasons and the intended statistical evaluation.

NOTE 1 The number of participating laboratories is less important than the number of independent data sets. A single laboratory might be able to provide several data sets, all obtained by independent procedures, calibrants and/or instruments.

The organizer should implement adequate quality control measures to ascertain the quality of the results delivered.

The producer should specify the form of reporting. The specification should include

- instructions on reporting of individual observations, averages, or both;
- the measurement units required for quantitative results;
- the number of significant digits required for quantitative results;
- where appropriate, the form of measurement uncertainty required;
- the nature and form of additional information required by the RM producer (such as measurement procedures and measurement standards used, dates and times of measurement, or run order).

Reporting can consist of individual results for each replicate measurement with or without uncertainty or one single result with stated uncertainty, which leads to different approaches to review and evaluation (see Annex A).

The form of reporting may also include preformatted reporting forms for participants. The reports should contain sufficient detail to check the technical validity of results, including information on traceability.

NOTE 2 When reports are submitted in spreadsheet form, unintentional alteration can often be prevented by the use of 'locking' or 'protection' facilities incorporated in the spreadsheet software.

The organizer should provide sufficient guidance for participating laboratories and/or operators to ensure the smooth implementation of the work. To be successful, the interlaboratory study should have a well-defined objective, be effectively designed and be efficiently organized with clear, concise guidelines with which all involved can readily comply. Participation, either as operator or as laboratory, in such a programme implies agreement to adhere to these guidelines.

NOTE 3 Additional guidance on the organization of multiple-method studies in one or more laboratories is given in Annex A, which is an extension of this clause.

9.4.3 Evaluation

9.4.3.1 Technical and statistical evaluation

Data sets should be inspected visually and graphically. Data submitted by each laboratory should be checked for completeness and any observed anomaly should be examined carefully for possible trivial (transmission error, misprint, etc.) and non-trivial reasons (drop-out, equipment failure, etc.). If transcription errors are suspected, the laboratory in question should be contacted to query the reported values, but the expected value should not be given at this time. If errors or failures are confirmed, the corresponding results should be corrected or rejected.

All results should be checked for evidence of technical errors based on the information on the measurement procedures provided by the study participants. The technical evaluation should lead to a set of technically valid data, i.e. data that each taken alone would be regarded as an unbiased estimate of the true value.

NOTE 1 The term technical errors refer to measurement results that can be excluded from the data set based on scientific evidence. The term does not refer to measurement data that is shown to be outlying from the data set based solely on statistical considerations.

NOTE 2 Inclusion of quality control materials, with known values, in such studies has been found useful to identify technical problems.

The pool of technically accepted data sets should be evaluated statistically, giving due consideration to evidence of between group differences (particularly between-method and between-laboratory differences), the underlying distribution of values, presence of clusters of results and potential outliers. Appropriate statistical methods for the data set and property to be certified should be selected.

Where the producer requires reporting of measurement uncertainty, the technical and statistical review should also consider the validity of any reported uncertainty information. Conclusions should take due account of the reported measurement uncertainties.

9.4.3.2 Assigned value and uncertainty

Value assignment should use appropriate statistical procedures. The procedure used should be valid for the particular data set.

NOTE Validation of statistical procedures can include evidence of a sound theoretical basis (usually by reference to appropriate literature), known performance under the expected conditions of use and assumptions or conditions which can be shown to apply to the data sufficiently for the purpose at hand.

Instruction on the use of two commonly used procedures, the mean and weighted mean, is given in A.2.4.

The uncertainty of characterization can be estimated either by using the uncertainty statements submitted by the laboratory or from the submitted data, ignoring the uncertainty statements made by the laboratory, or from a combination of both. More information is given in A.2.5.

9.4.4 Single-laboratory multi-method studies

In some cases, organisations have invested an exceptional amount of effort in method development, such that the metrological control of the measurement procedures approaches that of reference measurement procedures. In such cases, data sets from only a few of these measurement procedures, given that their measurement principles are sufficiently different, can be sufficient for characterization. Under these circumstances:

a) the RMP is likely to have access to the complete quality assurance and validation data, which should be taken into consideration for the technical evaluation;

b) the number of data sets is small. Therefore, more emphasis should be put on the assessment and proper treatment of measurement uncertainties. The evaluation should rely on the assessment and use of measurement uncertainties associated with each measurement procedure.

Where results agree within the claimed uncertainties, the weighted mean (A.2.4) and corresponding uncertainty may be used. Where apparently valid results do not agree well within the claimed uncertainty, one should carefully reconsider whether the metrological control of the measurement procedures is indeed sufficient for this approach. If this is confirmed, the effect of the excess dispersion of results should be allowed for in the certified value uncertainty.

NOTE Approaches that make allowances for excess dispersion include those of Mandel and Paule[38], Vangel and Ruhkin, Birge and others. Details can be found in Reference [34].

9.5 Characterization of an operationally defined measurand using a network of competent laboratories

9.5.1 Concept

This approach is applicable to the production of RMs certified for operationally defined measurands. As in this case the measurement procedure defines the measurand, demonstration of absence of a laboratory bias is often only possible by combining data from several laboratories. In addition, the defining procedure is often relatively imprecise and the only practical means of obtaining a small uncertainty is to average many results from different laboratories.

This approach is largely similar to that described in 9.4, with the exception that all laboratories apply the same procedure.

The assumption is again that a number of laboratories exist that can perform the measurement in question equally well. The approach aims at randomization of all influence factors within the limits set by the measurement procedure.

9.5.2 Study setup

A well-described measurement procedure should be chosen. This should be a published standard method, ideally an internationally agreed procedure (e.g. ISO, ASTM, AOAC or IFCC). Participants should be instructed to follow the procedure exactly, allowing only those variations that are permitted within the procedure.

NOTE 1 Any modification of such a procedure agreed by all participants (e.g. tighter specifications for some parameters) results in principle in a modified procedure and the measurand's identity is then defined by reference to the modified procedure.

NOTE 2 Preliminary studies can show unintended departures from the standard procedure, which can be corrected before proceeding to characterization in order to ensure adherence to the standard procedure.

NOTE 3 Quality control samples can also be used in this case to demonstrate that a particular instrument fulfils all specifications.

9.5.3 Evaluation

In the case of operationally defined measurands, the defining procedure is (by definition) unbiased and it is then necessary only to consider possible laboratory bias and within-laboratory effects in an uncertainty evaluation.

The approaches described in A.2 apply for the evaluation of results.

9.6 Purity

9.6.1 General

Pure substances constitute the primary measurement standard and ultimate source of higher-order metrological traceability for most traceability chains in chemistry, thermometry and calorimetry in general and for the certification of solution and matrix reference materials in particular. The adjective "pure" refers to an idealised situation: no substance is 100 % pure, there will always be impurities present at some level. The appropriate certification of substances for purity is thus an essential cornerstone of traceability in chemical measurement.

Pure substances are an important class of CRMs in their own right. They are used by laboratories either to disseminate higher order traceability to calibration standards used in measurement procedures, or in the certification and production of other CRMs, such as solutions or gas mixtures.

The purity of substances can either be determined directly (by measuring the amount of the substance in question) or indirectly by subtracting the mass or mole fractions of all impurities from 100 %.

When characterizing the purity of a material, the identity of the material should additionally be confirmed.

9.6.2 Direct determination of purity

In some cases, the mass or mole fraction of the substance in question can be determined directly. Suitable methods can include coulometry, titrimetry and calorimetry (freezing point depression). In the case of organic analytes the use of the technique of quantitative NMR for the direct certification of the purity of reference materials is increasingly being implemented[39],[40].

Methods requiring calibration with the substance in question (e.g. HPLC, GC, ICP-MS or AAS) can in principle be used for purity assignment, but they are secondary measurement procedures. Since they require a standard of known purity, the application of such methods for direct determination of purity is often limited to the assignment of values to working standards.

The purity determined by the procedure used is adopted as the assigned value and the uncertainty of the purity determination is adopted as u_{char}.

Confirmation of these values by independent measurements is highly recommended.

9.6.3 Indirect determination of purity

Purity can be determined by difference, using a set of orthogonal analytical techniques capable of detecting and quantifying all the major classes of impurities in the material, as follows:

a) a suitable range of possible impurities are investigated, often including residual organic solvents, water, inorganic and organic impurities. The types of impurities to be investigated are often informed by the manufacturing process for the substance;

b) the amount of each of the possible impurities is determined in the substance to be certified;

c) the purity of the main component is computed by difference.

The measurements necessary to determine the impurities can be challenging, since most impurities will be close to the detection and/or determination limits or can be difficult to resolve in the case of impurities closely related in structure to the main component. Furthermore, different measurement methods can give results in non-compatible units (mass fractions for volatiles; mole fractions for total impurities by calorimetry), which makes the combination of such results in a strict metrological sense impossible if the structures of the impurities are not known. Quantification of each impurity against specific calibrants is ideal, but can be impractical or impossible if sufficient resources or appropriate reference materials are not available. In this case, appropriate allowance has to be made for the uncertainty introduced as a result of assumptions regarding the identity and response factors of individual impurities[41].

NOTE The ICH harmonized tripartite guideline Q3A "Impurities in new drug substances"[42] requires that impurities above 0,05 % (depending on the daily uptake of the substance) be identified.

Although high relative uncertainties can be obtained for the quantification of individual impurities, provided the absolute level is small, the contribution to the uncertainty of the final value for the main component is usually low.

The model for the certified value y_{char} of the amount of substance or mass fraction of the main component y as a function of k, impurities with amount of substance or mass fractions w_i is given by Formula (14):

$$y_{\text{char}} = 1 - \sum w_i \tag{14}$$

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

$$u_{\text{char}}^2 = \sum u^2(w_i) \tag{15}$$

where $u(w_i)$ is the standard uncertainty in w_i. It frequently happens that some of the impurity amount of substance fractions or mass fractions w_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 procedure. Where a value for an impurity is below the detection limit, the value is sometimes set to zero and other times another value is assigned, often related to the limit of detection, with an associated uncertainty.

The evaluation of the uncertainties can also be complicated by the proximity of physical limits (amount of substance and mass fractions are only defined between 0 and 1), which can create additional problems, including estimates for some contributing impurity classes that include nominally negative values (see also Reference [51]).

9.7 Identity

9.7.1 Materials certified based on provenance

A reference material may be characterized based on knowledge of the origin of the material, i.e. the provenance of the material.

To support characterization based on provenance, the RM producer should obtain documentary or other evidence of the origin of the material that shows an unbroken chain of evidence from origin to final packaging. The documentation should be maintained for the lifetime of the material.

RM producers should have procedures in place to ensure that handling of the material (including sampling, homogenization, packaging, storage, etc.) prevents contamination by other materials and does not change the response of typical test methods for which the material is intended.

Whenever possible, RM producers should undertake experimental verification (including measurement, expert inspection or qualitative testing, as appropriate) to confirm the identity assigned using provenance.

EXAMPLE DNA extracted from a bacterial culture grown from a single bacterium, which in turn has been isolated from a bank of reference strain[43], could be certified based on provenance, subject to confirmatory checks for contamination.

9.7.2 Materials certified for identity based on measurements

9.7.2.1 General

When characterizing the identity of a substance based on measurements, several aspects should be borne in mind, including:

a) Identity is usually not a measurement result, but a conclusion drawn based on measurement results from one or several methods. For example, chemical shifts and the heights of peaks in an NMR spectrum, or a combination of colour, melting point, molar mass, etc., can inform an assignment of identity. While measurement uncertainties can be assigned to the individual measurement results, combining them to give any numerical indication of uncertainty in the identity (for example, a probability that the assigned identity is correct) is not straightforward.

EXAMPLE 1 Identification using DNA sequencing illustrates the difference between uncertainties in identity and uncertainties associated with measurement results. The DNA sequence is a result of a sequence determination experiment, and the probability of base pair and other errors in the sequencing as well as the presence of mutation differences between different DNA molecules can in principle be estimated. However, identity of a biological species can often be established with considerable confidence at relatively low percentage of homology with a reference sequence. Individual sequencing errors therefore might not materially affect the assignment of identity.

b) A CRM certified for identity is in practice only useful if the error probability on the conclusion is negligible.

c) Slight heterogeneity and instability of the material does not necessarily change the conclusion of identity. The guiding principle for the assessment of homogeneity and stability is applicability of the material, i.e. whether it still allows unequivocal identification.

d) Different substances can share the same properties for the identification methods chosen. Information on the source of the raw material and on the processing steps of the material to be characterized is therefore vital for the certification of identity.

e) As with any material, the project planning should establish a clear definition of the need for identity information based on the intended use of the material.

EXAMPLE 2 For DNA, the intended use could require only a statement of species identity, a complete sequence, or additional information on the degree of methylation.

NOTE Identity is sometimes determined by expert judgement (e.g. for asbestos fibres or microbial species). However, this judgement is usually based on observations and comparison with specifications. Expert judgement based on observations falls within the scope of this clause.

9.7.2.2 Specification

Testing for identity of a material involves comparison of a set of measurement results on that material with specifications (for example, melting point range; percentage of homology with a reference DNA sequence) for these measurement results.

EXAMPLE An organic polymer material might be identified based on comparison with a reference infrared (IR) spectrum using the following criteria:

— all peak frequencies in the reference spectrum are matched within 3 cm^{-1};

— relative peak intensities match the reference spectrum within 5 % absorbance;

— no peaks in the reference spectrum are absent;

— all peaks present in the candidate RM spectrum are present in the reference spectrum.

Sources of specifications can include internationally recognized compendia (e.g. Pharmacopeia sources[44] and other collections of reference data[45]). Such information can change outside the control of the RM producer. RM producers should therefore clearly state the specifications used for the assignment of identity, either as a set of values or as a dated reference on the certificate to an external specification.

When compiling specifications, RM producers should compare various literature data, establish the range of reported values and establish and document specifications for each measurand reflecting the ranges and reliability of the information used. Preference should be given to reference data which have undergone peer review.

9.7.2.3 Characterization of identity by a combination of methods

This approach is especially suitable for defined chemical substances of a small to medium molecular mass.

A number of methods should be chosen that probe different properties of the candidate reference material. Frequently used methods include, for example, determination of melting point, molar mass, UV, IR, NMR and mass spectra. Together with information on the raw material and its processing steps and the sampling and transport to the RM producer, the collection of methods should be sufficient to establish the identity of the material beyond any reasonable doubt. If detailed published specifications (e.g. Pharmacopoeial criteria for identification) exist, the choice of methods may be restricted to those listed in these specifications.

NOTE 1 The nature and number of methods required to establish identity varies with the number of potentially similar products (e.g. there are more organic than inorganic substances) and the information on the origin and processing steps.

All test and measurement procedures used should be properly validated and the results should fulfil the requirements for traceability laid out in 9.2. Where available, appropriate control materials should be examined alongside the RM, during characterization.

The results of each of the tests and measurements made should be compared with the specification for the proposed substance. Published procedures for such comparisons should be followed, where available. Where no such prescribed procedures exist, measurement results should not differ from any of the specified values when taking the combined uncertainty of measurement and specified value into account.

If the results agree with the specification, identity is established with a negligible uncertainty.

NOTE 2 A judgement on whether the accumulated measurement and provenance information is sufficient to establish identity beyond reasonable doubt is somewhat subjective. RM producers are therefore strongly encouraged to establish a system of peer review.

9.8 Presence/absence

Presence/absence is an example of a quantitative measurement that is evaluated in a qualitative manner. Results above a predetermined threshold are classified as “presence”; results below are classified as “absence”.

NOTE 1 Many measurements are evaluated as present/absent but are never quantified. Even for these methods, however, there is usually a limit for the response to be regarded as indicating “presence”. Existence of such a limit indicates the quantitative nature of the measurement despite the qualitative evaluation.

Quantitative evaluation of the measurements is one solution to this problem. Although the measurement uncertainty is frequently high (which is the reason for the qualitative evaluation), this approach has the advantage of being conceptually simple. The simplest case is characterization of a material for the absence of a substance for which quantitative methods exist (e.g. a contaminant in a foodstuff). In this case, all measurements on the material should give results below the critical value for declaring a substance present and the certified value is stated as “< L_d”, with L_d being the limit of detection. If test results are not quantitative, all measurements should provide the result “absent” to certify a material for the absence of a certain substance and the reference value in this case is stated as “absent”. Also, the limit of detection of the measurement procedure should be given.

NOTE 2 The term “Limit of Detection” is used in its IUPAC definition as “smallest measure that can be detected with reasonable certainty for a given analytical procedure”[46]. This refers to the (true) concentration where one is reasonably certain to detect a substance if it is present and corresponds to CCβ (in European Commission Decision 2002/657EC [47]) and to the “minimum detectable value of the net state variable” (as defined in ISO 11843-1[48]). For a given procedure, this limit depends on the number of replicate measurements.

For CRMs, the uncertainty statement should state the confidence level for any upper limit given for the concentration.

If several measurement procedures are used, and the results all agree, the limit of detection of the most sensitive procedure may be used as the certified value.

EXAMPLE Three different measurement procedures give results stated as < 2 mg/kg, < 5 mg/kg and < 4 mg/kg. As all results agree, the certified value is set as < 2 mg/kg. For an example, see Reference [49].

NOTE 3 Use of different procedures and/or laboratories can help to avoid misinterpreting losses during the analytical procedure as absence. Furthermore, inclusion of information on the processing of the material can be used to support the statement of absence of a certain substance.

NOTE 4 It can be necessary to declare a substance to be present if the result is above a critical value, even if it is below the limit of quantification. See for further guidance ISO 11843-1[48].

9.9 Ordinal scales

Some properties are expressed on an ordinal scale, which usually places items in ordered classes. Examples are the Mohs hardness scale or skin irritation classified as no response/moderate redness/significant irritation/severe reaction. These scales are often defined by reference to a particular method of classification. The only possible characterization approach in that case is therefore characterization by several laboratories using the same method.

In many cases, an RM will only be useful if it is put into one class without any disputes. To achieve this, all technically accepted measurements by all participating laboratories should put the material into the same class.

If some results deviate and the deviation cannot be explained by technical errors, no class can reliably be assigned. It can, however, be useful to give the median and/or the mode of the technically valid results as an information value.

9.10 Qualitative properties

Materials can be characterized for qualitative properties such as colour, odour or shape[82]. In some cases, these properties can be quantified and are in practice often used in this quantified form. Examples are the shape parameters of particles or colour according to the Hunter system[50]. This transforms the problem to the characterization of a method-defined measurand as described above. For colour especially, characterization of the absorbance/reflectance spectrum may also be considered.

9.11 Characterization of non-certified values

According to the definition of “reference material” and “certified reference material”, only certified values need to be accompanied by a measurement uncertainty and a statement of metrological traceability. Apart from the certified values, non-certified values (named, for example “indicative values”, “information values” or “informative values”) may be assigned, which, however, cannot be used as a reference in a metrological traceability chain.

As there is no requirement for uncertainty and traceability (i.e. no requirement for comparability) of such values, a wider range of approaches can be used for value assignment, including use of literature data about typical properties, circumstantial data from single laboratories or pooled data from several laboratories. The closer the chosen characterization approach resembles an approach appropriate for certified values, the more reliable this assigned non-certified value will be.

It is recommended to give information on the origin of the non-certified value, and why it is not certified, to allow users to assess its fitness for purpose. While not required, traceability statements and statements of uncertainties increase the usefulness of these values.

10 Evaluating measurement uncertainty

10.1 Basis for evaluating the uncertainty of a property value of a CRM

The basic principles for evaluation of the uncertainty of the certified value in this document are the general principles set out in the GUM[4]. The GUM provides a procedure for combining different contributions to uncertainty, each expressed as standard uncertainties. A summary of the procedure is given in Annex D.

This detailed procedure is appropriate where a complete measurement model can be written and where all sources of uncertainty are associated with influence quantities in the model. In many cases in RM production, however, some or all of the major uncertainties are not associated with measured input quantities and a simplified model, such as that given in 10.2, is appropriate.

10.2 Basic model for a batch characterization

The value of a certified property in a single unit of an RM when delivered to the user can, in principle, be affected by the characterization process, by real variation between individual units (heterogeneity), change over time and changes during transportation and subsequent storage. The model used for evaluating the uncertainty associated with a certified value should allow for all of these effects where they are significant. A convenient simple model for this purpose is as follows:

$$x_{\text{CRM}} = y_{\text{char}} + \delta_{\text{hom}} + \delta_{\text{lts}} \tag{16}$$

where

x_{CRM} denotes the property value;

y_{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;

δ_{hom} denotes an error term due to heterogeneity which includes between-unit variation together with any necessary within-unit heterogeneity allowance;

δ_{lts} denotes an error term representing the stability effects under storage conditions.

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

Formula (16) provides a simple additive ‘measurement model’ to which the GUM principles can be readily applied. Assuming independence of the variables, the uncertainty associated with a property value of a CRM can be expressed as

$$u_{\text{CRM}} = \sqrt{u_{\text{char}}^2 + u_{\text{hom}}^2 + u_{\text{lts}}^2} \tag{17}$$

using the law of propagation of uncertainty in the GUM[4]. The evaluation of these uncertainty components is covered in this document in 7.11 (for u_{hom}) and 8.7 (u_{lts}). Additional guidance is given in Clause 9 (characterization) on the evaluation of the uncertainty u_{char} arising from the characterization of the property value.

NOTE 1 In some circumstances, it can be desirable to include additional terms to Formulae (16) and (17). For example, where transportation cannot be sufficiently well controlled to guarantee negligible change, a further effect δ_{trn} and associated uncertainty u_{trn} can be added to the model given here.

NOTE 2 The combined standard uncertainty associated with the property value of the CRM can be related to the period of validity of the material (see 8.6).

NOTE 3 Sometimes, the property value is a function of time, such as for reference materials certified for radioactive isotopes. In these cases, ISO 17034 requires due allowance to be made in the stated uncertainty for possible change in the value prior to use or, where the change with time can be predicted, that a means of correcting the certified value and its uncertainty for the expected change over time be provided.

10.3 Uncertainty sources

The characterization uncertainty u_{char} should consider all the relevant uncertainty sources encountered in the measurement procedure(s) used for characterization. Both the GUM[4] and Reference [51] list common uncertainty sources. Often, the measurement procedures have already been evaluated in terms of measurement uncertainty, and the resulting uncertainty evaluation can be applied for the evaluation of the uncertainty of a property value of a CRM.

Often, uncertainty models of measurement procedures contain aggregated components, i.e. uncertainty components which depend on several others. These aggregated components can lead to covariances (see 10.1), even if these do not appear when the measurement procedure 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 the GUM (ISO/IEC Guide 98-3:2008, Annex F [4]), some further guidance is given on how to evaluate the resulting covariances.

Any change in a particular measurement procedure should be accompanied by a review of the uncertainty model.

10.4 Coverage intervals and factors

Following calculation of the standard uncertainty u_{CRM} in Formula (17), it is usually necessary to calculate an expanded uncertainty U_{CRM} in order to provide an interval $x_{\mathrm{CRM}} \pm U_{\mathrm{CRM}}$ including a large fraction of the values that could reasonably be attributed to the property being certified. The expanded uncertainty is calculated using Formula (18):

$$U_{\mathrm{CRM}} = ku_{\mathrm{CRM}} \tag{18}$$

where k is a coverage factor.

The coverage factor used is determined on the basis of the distribution function assumed and the coverage probability required. In many cases the distribution can be assumed to be approximately normal and the required coverage probability is 95 %; this leads to a coverage factor $k = 2$.

When the effective degrees of freedom associated with u_{CRM} is low (less than 10), the Student's t-distribution should be used instead of assigning a coverage factor of 2.

In cases where the assigned distribution of the property value is considered to be asymmetric, such as in the case of the result of a count following the Poisson distribution, a confidence interval (of a specific probability, i.e. 95 %) should be stated rather than the expanded uncertainty and a coverage factor.

NOTE Coverage intervals for asymmetric distributions will usually be asymmetric.

Annex A
(informative)

Design and evaluation of studies for the characterization of a method-independent measurand using two or more methods of demonstrable accuracy in one or more competent laboratories

A.1 Study design

A.1.1 Selection of measurement procedures

When selecting measurement procedures, variation of among others the following aspects should be considered:

— sample preparation, for example grinding/milling, extraction or clean-up steps;

— sample introduction and/or separation, for example using LC or GC;

— quantification principles, for example molecular or atomic absorption, mass spectrometry, flame ionization or fluorescence;

— calibration procedures, unless one approach has clear advantages, because of its metrological rigour or because of achieving lower measurement uncertainties.

In many cases, variation of all aspects will be impossible. In these cases, the maximum possible variation should be sought. For example, if gas chromatography is the only available separation technique, then the study should at least aim to include different injection techniques, different columns and temperature programs and quantification by different detectors.

The RM producer should require that all measurement procedures used in the campaign are properly validated and that a reasonable estimate of the measurement uncertainty can be provided.

To allow laboratories the free choice of measurement procedures while ensuring the necessary variation, the RM producer should obtain information about the measurement procedures applied by the participants before the start of the study to obtain the necessary range of procedures. If required, a targeted search for laboratories offering specific methods should be performed to avoid receiving results obtained by only one method[83].

A.1.2 Choice of calibration standards

An important decision is whether all laboratories should use the same calibrator or whether laboratories should be given free choice of the calibrator. Using a single calibrator reduces variation caused by different calibrators from different suppliers. On the other hand, any bias in this single calibrator will translate into the same bias in the certified values. Therefore, use of a single calibrator requires very careful characterization of this calibrator, requiring in many cases independent confirmation of purity or composition. Allowing laboratories the free choice of calibrator is logistically easier, does not require tests for independent confirmation and allows laboratories to apply their procedures unchanged, but means that some way of checking the appropriateness of the calibration should be included in the study. As a general guideline:

— for well-established measurements, where experience shows that the quality of available calibration standards is sufficient, giving laboratories the free choice of standards is usually preferable;

- in cases where there is significant doubt about the quality of standards on the market, the efforts needed to characterize a common standard are often justified.

In some cases, different producers of commercial standards obtain their pure material from the same company. It is therefore useful to establish whether the original sources differed.

A.1.3 Selection of laboratories

Laboratories should be selected based on demonstrated competence. Appropriate evidence for the demonstration of competence may include the following:

- results from proficiency tests;
- results on independent CRMs (possibly distributed as quality control materials together with the candidate CRM);
- method validation data;
- a full and credible uncertainty budget;
- previous participation in other RM certification campaigns for the same measurand; and
- third party assessment of conformance with ISO/IEC 17025[52] or other relevant standards for the determination of the measurand in question.

The performance of externally assessed laboratories can differ in the same way as other competent laboratories. It is therefore prudent to obtain information on performance in addition to evidence of third party assessment of conformance with ISO/IEC 17025.

The RMP should ensure that the measurements in each laboratory are performed in accordance with ISO/IEC 17025. In particular, the provisions of ISO/IEC 17025 regarding competence of staff, calibration of equipment and authorization (official release for use[52]) of methods and results should be met.

In the absence of independent assessment, information on the extent of the laboratory's quality systems should be obtained.

NOTE 1 Potential gaps in fulfilling the requirements of the respective standard can sometimes be filled by the RM producer (e.g. an RM producer can archive the laboratories' raw data, if they do not have an archiving system).

NOTE 2 Obtaining calibration certificates for each and every instrument is in many cases impractical. Agreement of results on quality control samples can be used as demonstration of sufficient calibration of all relevant input factors.

A.1.4 Number of independent data sets

The number of participating laboratories is less important than the number of independent data sets. A single laboratory might be able to provide several data sets, all obtained by independent measurement procedures. The remainder of the discussion focuses on data sets, regardless of whether each data set was provided by a different laboratory, some laboratories provide more than one data set or all data sets are provided by the same laboratory.

Complete independence of results is difficult to achieve if measurements are performed in a single laboratory. The RM producer should critically review the variation of all critical steps as outlined in A.1.1 to check whether sufficient method variability is present to demonstrate absence of bias for each individual step. This includes checking whether the same critical chemicals, equipment, calibrators, etc., were used by the various measurement procedures. The organizer should also remind the laboratories at the onset of the study not to censor data, i.e. not to suppress/change data from one procedure without notification after a crosscheck of results between the various procedures.

The RM producer should set a documented minimum number of technically valid results for which value assignment will be considered. The number of data sets should be large enough to provide an adequately small uncertainty in the estimated value after allowing for the possibility of failure to report, exclusion of results for technical reasons and the statistical evaluation intended.

For interlaboratory studies using a network of testing laboratories, the characterization should include five or more participants providing technically valid data.

NOTE A characterization uncertainty less than one third of the interlaboratory reproducibility standard deviation requires at least nine participants unless laboratories are selected for exceptional performance.

The following considerations affect the number of data sets required in order to achieve the desired uncertainty for certified values.

— **Uncertainty required**: The uncertainty of assigned values usually decreases with the number of technically valid data sets, requiring more data sets for smaller uncertainties.

— **Technical difficulty**: The less well established or the more technically challenging a measurement is, the larger the between-data set variation can be expected to be.

— **Likelihood of technically invalid results**: Even experienced laboratories can deliver technically invalid results which cannot be used for certification. Such results are more likely for unfamiliar materials, new or modified measurement procedures, challenging measurements or unusual reporting requirements. The number of participants or (for single laboratory studies) independent measurements should be increased where the risk of technical errors is higher.

— **Reproducibility/repeatability ratio**: Where between-data set variation is known to be the primary source of variation, preference should be given to a larger number of data sets rather than higher number of replicates for each data set.

— **Statistical evaluation**: Different data treatment methods can require larger numbers of observations to provide good numerical stability and/or achieve sufficiently small uncertainty.

— **Validation results from a CRM**: Results of consistent high quality obtained from one or more similar CRMs, used for quality assurance at each laboratory, serve to demonstrate the validity of all data sets and to provide information on bias among laboratories and measurement procedures.

A.1.5 Number of units and replicate determinations

The number of units of the candidate CRM sent to each participant, the number of replicate determinations performed by each participant and the condition of these determinations are determined by practical as well as evaluation considerations.

— If the variation between individual units of the RM is large, single measurements on several different units are preferable to several replicate measurements on a single unit. If contamination, breakage or heterogeneity are not an issue, sending a single unit is sufficient.

— In the absence of reliable uncertainty evaluations by the participants, requesting measurements under conditions of intermediate precision can provide an indication of the reliability of the participants' results or a check on the reliability of the participants' uncertainty estimates.

— If each laboratory receives more than one unit, more measurements can be made on the remaining unit(s) in case of breakage of one unit, which eliminates the need for a new dispatch.

A.1.6 Quality control materials

Inclusion of additional samples for quality control has been found to be highly beneficial. Results on these samples can identify technical problems and aid the technical evaluation.

— RMs, in particular natural matrix RMs and quality control (QC) materials, may be used to demonstrate the validity of the measurement result when measured alongside the unknown material to be characterized.

- Spiked materials, spiked blanks, etc., may be used to check parts of the measurement procedure or to assist in the process of assigning values to a material.
- Blank matrix materials, blank extracts, etc., may be used to demonstrate that the measurement procedure 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 together with the uncertainty of the correction factor.

NOTE Sometimes CRMs used for quality control in an interlaboratory study are supplied without the original label to avoid identification. However, as the number of reliable CRMs is limited, experienced laboratories can recognize the material from the visual appearance and/or the values.

A.1.7 Instructions for participants

Guidelines to participants should contain:

a) a clear outline of the goal of the study;

b) instructions to refrain from comparing results with other participants, including the reasons for discouraging collusion (that is, cooperative exchange of information);

c) the number of units to be tested;

d) the number of replicate determinations to be performed;

e) any restrictions or specific details of measurement procedures to be used; for example, any need for prior drying and moisture correction;

f) the minimum test portion size;

g) requirements with respect to quality and traceability of the measurement results;

h) the time schedule (distribution of samples, delivery of results);

i) the mode of dispatch;

j) instructions for intermediate storage of samples;

k) specific instructions for sample treatment, if applicable;

l) instructions on quality control measures to identify potential bias; and

m) information on the producer's policy on identification of laboratories and use of data; for example, whether laboratories will be identified, whether results will be identified with a particular laboratory and whether the results may be used for purposes other than the characterization study.

NOTE 1 A meeting with the laboratories/groups involved (prior to distributing the samples and performing the measurements) can help all parties involved to align all actions to be carried out during the collaborative study, and to discuss possible problems and/or pitfalls.

In an interlaboratory study the RM producer should take reasonable steps to prevent collusion between laboratories, including but not limited to b) above.

NOTE 2 Different labelling of materials for each participant can make it harder for participants to compare results.

A.1.8 Reporting

The use of preformatted reporting forms can be useful as it has the advantage of structuring the report and (if transmitted electronically) allows copying of the results, which can reduce transcription errors in the RM producer's collation of results. Disadvantages of preformatted reporting forms are that they often force laboratories to depart from their usual reporting practices, which can lead to transcription errors. If reports are submitted electronically, the requirements of ISO 17034 on the integrity of electronic records, especially of reports of test results, should be adhered to.

NOTE When reports are submitted in spreadsheet form, unintentional alteration can often be prevented by the use of 'locking' or 'protection' facilities incorporated in the spreadsheet software.

Laboratories can be requested to report individual results (not only averages over all samples), regardless of whether an uncertainty statement is reported or not, although reporting of an average and an expanded uncertainty and its coverage factor can be sufficient.

Where there is an option for correction of a known procedural bias, such as extraction recovery, the RM producer should state clearly whether results should be corrected or not. The RM producer may also require participants to report bias checks (e.g. from spikes) and use these to correct results for detected bias. Where a correction is applied by the participant, any reported uncertainty should include the uncertainty associated with the correction.

Participants should be instructed on how to report results near detection limits, if such results are likely to occur. Results reported as “less than” make statistical evaluations more difficult. On the other hand, reporting of results near detection limits contradicts many laboratories' quality procedures. Where results near detection limits are likely, RM producers should either require laboratories to report the observed (including negative) results instead of, or in addition to, their normal reporting, or should adopt statistical procedures that allow for “left-censored” results such as “less than” statements or results restricted to values above zero.

Instructions to participants should specify the measurement units and number of significant digits to be reported for quantitative results.

It is recommended that an outline of the measurement procedure used is reported in sufficient detail to permit an understanding of all stages in the measurement process (e.g. in chemical analysis, the digestion/extraction of the sample and separation of the analytes of interest, clean-up, and quantification). Participants should be requested to give literature references, where applicable.

A.2 Evaluation

A.2.1 General considerations for evaluation

In the course of evaluation, anomalies can arise that require communication with the participant concerned. The producer may contact participants to assist in the investigation of anomalies at any stage of the evaluation process. If a participant is contacted, it is recommended that initial contact should not specify the nature of the anomaly (for example, the direction of deviation of the results); rather, the participant should initially be invited to investigate and report any errors discovered.

If data sets from more than one measurement procedure are provided by a single laboratory, initial inspection should consider the data sets individually (that is, as if independently reported by different laboratories). The RM producer should nonetheless take account of all data sets submitted by a laboratory when drawing conclusions about which of the data sets to retain when anomalies are found.

A.2.2 Initial screening

A.2.2.1 General

Information from each participant should be examined on receipt or as soon as practicable after receipt.

Initial examination of individual participant results should check for evidence of basic reporting or procedural errors such as missing data (including any requested information that is absent); incorrect numbers of replicates; inappropriate conditions of measurement (for example, repeatability versus reproducibility conditions); incorrect identification of test items (e.g. through accidental mislabelling) and incorrectly reported units of measurement. Apparent errors at this stage should, where possible, be referred promptly to the participant for checking and possible correction (see A.2.1).

Unexpectedly high or low results or uncertainties can also be apparent on receipt and may be referred to the participant for checking at this stage.

A.2.2.2 Technical evaluation

Further technical examination to identify potential problems may include (but is not limited to) grouping results by techniques (measurement procedure and principle, sample pre-treatment methods, etc.) or the calibration technique used. In addition, comparison of the expanded uncertainties with the confidence interval of the mean of the submitted results gives an indication of whether the stated uncertainty is realistic, as the expanded uncertainty should be at least as large as the confidence interval.

Technical examination of individual participant results should, as far as possible, check for evidence of errors in procedure such as: reported use of an inappropriate measurement procedure or pre-treatment procedure; inappropriate calibration; inappropriate conditions of measurement (for example, repeatability versus reproducibility conditions) and incorrect units of measurement. Where measurement uncertainty is reported, the examination should check for unusually high or low uncertainty compared with typical expectations and, where uncertainty budgets are available, for omission of significant contributions to uncertainty or inappropriate uncertainty evaluation methods.

Technically invalid results should be removed from the data set or corrected, by repeating the measurement, if possible and necessary.

NOTE 1 A technically invalid result is not necessarily an outlier nor is every outlier necessarily technically invalid. A result can fall well within the range of valid results, even when it is evident that the conditions, under which the result was obtained, were not in good order. Conversely, a result deviating significantly from all other results can be the only technically valid result of the data set.

NOTE 2 Appropriate choice of statistical procedures for value assignment can allow useful value assignment even when reported uncertainties show evidence of, for example, under-estimation. Reference [34] provides further guidance on such procedures.

A.2.3 Statistical evaluation

A.2.3.1 General principles for statistical evaluation

Characterization by a collaborative study or by multiple measurement procedures aims at randomization of bias between data sets. Statistical evaluation typically assumes that the true value of a measurand corresponds to the true value of a population parameter, usually the population mean.

Different procedures intended to estimate the value of a particular measurand – whether the measurand is operationally defined or not – can be systematically biased as well as showing laboratory specific bias per data set. The possibility of between-method differences should normally be considered in evaluating measurement uncertainty.

Where measurement uncertainty is reported, statistical evaluation should check for unusually high or low measurement uncertainty, uncertainty/location anomalies such as results far from any central estimate compared with their reported uncertainty, and any evidence of generally inadequate uncertainty evaluation (for example, greater dispersion than accounted for by reported uncertainties). Anomalies related to reported measurement uncertainties should be resolved where possible, for example by referral to participants for checking and possible correction.

A.2.3.2 Distributions

Finding appropriate estimators for the expected value is closely linked to the (either assumed or determined) underlying distribution of values.

Determining a probability density function from a data set requires significantly more data than can practically be obtained, whereas checking for consistency with an assumed distribution and calculating parameters from it (e.g. average of a normal distribution) is possible with the number of data sets usually available in characterization studies.

The RM producer should therefore check whether there is evidence of deviation from the assumed distribution using, for example, visual methods (histograms, kernel density plots and normal probability or, more generally, Q-Q plots) or statistical checks for departure from particular distributions, including tests for normality or determination of skewness and kurtosis. An approximately normal distribution of data sets is often observed for results well above the limit of quantification; other distributions include Poisson distributions (e.g. microbe counts) or a Weibull distribution (e.g. mechanical failure of ceramics). The selected distribution should be in agreement with the reported data as well as with the theoretical and historical knowledge of the measurement in question. If these differ significantly, no value should be assigned unless technical reasons for the unexpected distribution can be given.

EXAMPLE Theoretical considerations as well as a multitude of interlaboratory comparisons indicate that results for trace elements in soil follow normal distributions (unless the level is close to the limit of quantification). Deviation of the observed data from normal distribution (e.g. tailing) indicates technical problems. If it is not known whether the problem lies with the majority or with the tails, no value is assigned.

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. There should be a technical basis for such a transformation, as deviation of the results from the expected distribution may indicate technical problems.

A.2.3.3 Outliers

Outlying 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 may be identified by, for example, appropriate outlier tests for outlying means and variances, graphical inspection of raw data, and use of Mandel's h and k statistics[24].

Outlying observations or mean values should not be removed solely on the grounds of a statistical outlier test, but may be removed if there is a technical reason to do so.

NOTE 1 Technical reasons include inadequate calibration, inadequate measurement procedures, use of inadequate reagents, failure to account for interferences and deviation from the certified value of an independent quality control material.

NOTE 2 Outlier tests usually ignore measurement uncertainties. An outlying data set can agree with the other data when the respective uncertainties are taken into account.

NOTE 3 Data points, with unusually large uncertainty, can be removed on technical grounds, if they agree with the other results within their uncertainties, and if the uncertainty is so large compared with the other reported uncertainties that it indicates a technical problem. They can also be retained, because they agree with the rest of the results.

NOTE 4 Data sets that show a high outlying variance can indicate a lack of method repeatability or lack of intermediate precision, which can justify rejection on technical grounds.

NOTE 5 Different measurement procedures may differ in precision and it can be important to retain data sets for a relatively imprecise procedure in order to retain a representative collection of procedures.

NOTE 6 A special case of extreme variance is a set of results having zero variance. This can arise when laboratories report too few significant digits. This can be prevented by appropriate instructions to participants (A.1.7) or corrected by reference to the participant. Use of the mean or median of such a group of identical data can also be valid, for example where between-group effects dictate the use of group means.

A.2.3.4 Robust statistics

Robust statistics provide a large collection of statistical methods that explicitly allow for the presence of outlying values in an otherwise approximately normally distributed data set. Typically, robust methods assign weights that decrease with the distance from the main body of the data. Robust methods exist for estimating the mean and standard deviation for simple outlier-contaminated data sets, as well as for many other parameters of interest. Robust methods also exist for other situations including, for example, sets of reported values with appreciably different uncertainties, or the analysis of data consisting of multiple replicates for each data set.

Robust estimators provide high resistance to the influence of extreme outlying values. For symmetric 'heavy-tailed' distributions (that include a larger proportion of values far from the mean than would be expected from a normal distribution), robust statistics typically provide unbiased estimates of the mean with lower variance than the simple arithmetic mean. This approach results in an unbiased estimate with smaller uncertainty than the arithmetic mean would give in the same circumstances. For data with extreme outliers, robust statistics can be a very considerable improvement over the mean. A short summary of useful robust statistical methods, together with conditions for their use in RM characterization and references to more detailed descriptions, is given in B.5.

NOTE The tolerance of robust statistics to extreme values additionally makes them useful for identifying outliers in larger data sets containing several extreme values.

A.2.3.5 Grouping ("clustering")

Statistical evaluation should check for the occurrence of grouping of results, for example along measurement procedure, calibrants, reagents or regions.

a) If the difference between means for different groups is statistically significant and is too large to permit a sufficiently small uncertainty for the intended use of the material, then no single property value can be assigned. Where grouping is along reagents/calibrants, the technical evaluation should check whether all of them are appropriate. Where the grouping is along measurement procedure, the producer may provide an assigned value for each measurement procedure.

b) If there are significant differences and the difference between the means of these groups is relatively small, one single value may be assigned. An additional uncertainty term accounting for the between-group variation should then be added to the uncertainty of characterization. There are several approaches to this estimation problem described in the literature[53],[54].

c) If the difference between these clusters is large and there is no correlation of these clusters with measurement procedures or other technical explanation for the differences, no value can be assigned. A larger pool of results can be necessary to overcome the relatively poor agreement of measurement procedures available.

NOTE Visible grouping can arise from other causes, including chance (especially in small data sets), regional differences in application of a procedure or use of different calibration standards.

A.2.4 Assigned value (weighted/unweighted mean)

Where the data set means follow an approximately normal distribution and no weighting is applied, the unweighted arithmetic mean of the p data set means y_i is applied as assigned value y_{char}.

$$y_{\text{char}} = \frac{\sum y_i}{p} \tag{A.1}$$

The mean value of individual results may also be adopted as the assigned value where differences between data set means are insignificant compared with the effects of variation within each data set.

A weighted mean is usually calculated using the general form of Formula (A.2):

$$y_{\text{char}} = \sum_{i=1}^{p} w_i x_i \Big/ \sum_{i=1}^{p} w_i \tag{A.2}$$

Where w_i is the weight applied to each data set mean (or, where a value and uncertainty are provided, to a single value) x_i.

The simplest choice of weights w_i is given by Formula (A.3):

$$w_i = 1/u_i^2 \tag{A.3}$$

where u_i is the reported standard uncertainty for the value x_i. This scheme should be used only when the reported uncertainties can be shown to be reliable and where it is clear that differences between groups can be wholly accounted for by the reported uncertainties. Where the uncertainties do not account fully for the observed dispersion of values x_i, alternative schemes should be used. Alternative approaches to weighting that are less susceptible to the effects of unreliable uncertainty evaluation are described in, for example, References [33] and [34]. Where reliable weights cannot be ascribed, the simple mean is often a useful conservative alternative.

If the results do not follow a normal distribution but can be transformed into normally distributed data, the data can be evaluated according to the following steps: transform the raw data; apply the calculations as in Formula (A.1); calculate the assigned value and confidence interval; apply the reverse transform to the assigned value and confidence interval. If a standard uncertainty is also required, either refer to statistical texts on the variance of distributions, or use the GUM to obtain it, bearing in mind that first-order error propagation can fail badly for relative uncertainties over 15 % and higher-order terms can be needed.

If the results do not follow a normal distribution and cannot be transformed into normally distributed data, a statistically sound approach that is consistent with the observed distribution should be applied.

A.2.5 Assigned uncertainty

A.2.5.1 Use of analysis of variance for uncertainty evaluation

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 or the between-laboratory standard deviation. Otherwise, the mean of means may be computed for these strategies instead.

A.2.5.2 Uncertainty-based evaluation

It is theoretically possible to combine the results, including their uncertainties, into a single value (the property value) and a combined standard uncertainty.

Approaches described include weighting results by uncertainties[35], determining detailed expressions for the uncertainties[34], least squares fitting (for example Reference [36]) or splitting uncertainties into common and individual parts[37].

NOTE These approaches are possible in theory, but are often difficult to implement in practice and have so far rarely been used.

A.2.5.3 Evaluation without the laboratories' uncertainties

Where Formula (A.1) was used to calculate the certified value, where the data set means follow an approximately normal distribution and no weighting is applied, the standard deviation of the mean of the p data set means y_i can be applied as u_{char}.

$$u_{\text{char}} = \frac{s(y)}{\sqrt{p}} = \frac{1}{\sqrt{p}} \cdot \sqrt{\frac{\sum (y_i - y_{\text{char}})^2}{p-1}} \tag{A.4}$$

where $s(y)$ denotes the standard deviation of the p data set mean values.

If the results do not follow a normal distribution and cannot be transformed into normally distributed data, a metrologically and statistically sound approach that is consistent with the observed distribution should be applied.

NOTE Formula (A.4) only covers the between-laboratory parts of the uncertainty. The uncertainty of the assigned value of the calibrant can be added if many or all data sets were obtained by calibration with the same material, the uncertainty of which is significant compared with the result of Formula (15).

A.3 Use of collaborative studies for multiple purposes

A.3.1 There can be different purposes for interlaboratory comparisons, including value assignment of reference materials, evaluation of the performance of laboratories and evaluation of the performance characteristics of a measurement procedure. In general, a particular study is best used for only one of these purposes; to do otherwise can compromise one objective in favour of another, or confuse the purpose of the study for the participants. Nonetheless, it can be useful to consider combining characterization studies for RMs with other studies to save costs, providing that due care is taken to avoid the principal disadvantages and that certain conditions are met. A.3.2 provides guidance on the principal disadvantages; A.3.3 provides conditions for combination of such studies with RM characterization.

A.3.2 Potential incompatibilities or conflicts when combining PT or method performance studies with characterization studies include the following.

a) PT studies aim to assess the competence of a laboratory whereas participation in characterization studies is restricted to laboratories of demonstrated competence. A combination of a PT study with a characterization study might therefore assume the very fact of demonstrated competence for the purpose of characterization that should be assessed in the PT part.

b) The requirements of proficiency testing do not normally require extensive replication and rarely permit specification of equal replication by all participants.

c) PT results are typically treated as confidential and access to details of experimental methods cannot be guaranteed.

d) Participants often pay for participation in a PT study and the motivation for participation can sometimes be commercially driven. Participants are therefore often unwilling to adhere to the detailed study setup, requirements for traceability and quality assurance, calibration, provision of uncertainties, etc., required in a characterization study.

e) The results of PT can be used for demonstration of competence to potential customers, accreditation bodies and others. Participants therefore can be less willing to discuss freely and to admit to technical problems required in a characterization study for the technical evaluation of results.

f) PT studies usually assess laboratory performance in comparison with other laboratories or with an externally set criterion, whereas a characterization study aims at getting a good estimate of the true value and its uncertainty. This leads to different approaches in evaluation, especially in dealing with extreme values/outliers.

g) PT studies are typically open to any laboratory willing to participate. RM producers therefore often have little control over the measurement procedures applied by the participants, which can result in receiving results from only one procedure even if other procedures exist.

h) While feedback on the results for the purpose of checking anomalies cannot commence before a PT study closes, this information is required for following up apparent anomalies of results in a characterization study.

i) The evaluation of a PT study also focuses very much on statistics, whereas characterization studies place more emphasis on the technical evaluation.

j) Method performance studies evaluate the performance of a measurement procedure immediately after development, so there is little prior information on typical performance.

k) Method performance studies typically apply to only one measurement procedure and are therefore not suited for assigning values to method-independent measurands.

A.3.3 Because of these potential conflicts, where a reference material is to be certified using data collected from participants in a PT scheme the RMP should:

a) decide, before the start of a study, which subset of data from specified laboratories will be used for value assignment;

b) demonstrate the competence of these laboratories independently of this study, e.g. from performance in previous rounds of the scheme or by other means;

c) organize the study for these selected laboratories according to the criteria described in A.1.3;

d) evaluate the results of these laboratories according to the criteria laid out in A.2;

e) notify proposed participants that results may be used in a certification study and obtain permission to access technical detail.

Annex B
(informative)

Statistical approaches

B.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 can be "missing" for a variety of reasons. Expressions have been developed to account for these missing data[55],[56] and are recommended for treating incomplete data sets; however, the more incomplete the data set becomes, the poorer the quality of the estimates.

One-way ANOVA applied to reference material studies is usually based on the statistical model

$$x_{ij} = \mu + \delta_i + \varepsilon_{ij} \tag{B.1}$$

where

x_{ij} is observation j in group i;

μ is the true mean value of the population of possible results (from which the observed results x_{ij} are assumed to arise);

δ_i is the (true) deviation of the group i mean from μ;

ε_{ij} is the random error for observation j in group i.

δ_i is assumed to arise from a population (usually normally distributed) with mean zero and standard deviation σ_{between} and ε_{ij} is assumed to arise from a population (again usually normally distributed) with mean zero and standard deviation σ_{within}. Note that the true standard deviation for each group is assumed to be the same for all groups. These two standard deviations account for the spread of data observed; the purpose of ANOVA as applied here is to estimate these two standard deviations.

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 hierarchical levels in the analysis of variance[55]. The so-called within- and between-group mean squares M_{within} and M_{between}, respectively, as obtained from a spreadsheet or statistical software program, can be converted into the estimated variances following Formulae B.2 to B.4:

$$s_{\text{within}}^2 = M_{\text{within}} \tag{B.2}$$

$$s_{\text{between}}^2 = \frac{M_{\text{between}} - M_{\text{within}}}{n_0} \tag{B.3}$$

where

$$n_0 = \frac{1}{a-1}\left[\sum_{i=1}^{a} n_i - \frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i}\right] \tag{B.4}$$

and s_{between} and s_{within} are the estimates of σ_{between} and σ_{within}, respectively.

When there are no missing data in a study planned to contain n observations per group, n_0 becomes equal to n.

Where necessary, the grand mean $\bar{\bar{x}}$ may also be estimated from Formula B.5:

$$\bar{\bar{x}} = \frac{1}{\sum_{i=1}^{a} n_i} \sum_{i=1}^{a} \sum_{j=1}^{n_i} x_{ij} \tag{B.5}$$

In the absence of any between-group effect, s_{between} is expected to be (close to) zero. If, for experimental reasons, a negative value is obtained for s_{between}^2 , then it should be set to zero.

NOTE When the initial estimate of s_{between}^2 is negative, $s_{\text{between}} = 0$ and $s_{\text{within}} = s(x)$, where $s(x)$ is the standard deviation of all the observations x_{ij}, form the maximum likelihood estimates of the two parameters (identical to the restricted maximum likelihood estimate). This is also a statistically valid procedure.

B.2 Two factor ANOVA for nested designs

This model may be used when the results of a collaborative study are used to confirm the homogeneity of the material, as well as to characterize it, and where the repeatability is constant across groups. The experimental scheme is illustrated in Figure B.1 for the particular case of an inter-laboratory study. When a campaign consists of different measurement procedures, the layout for the campaign is essentially the same.

The results can be described by the model Formula (B.6):

$$x_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk} \tag{B.6}$$

where

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

μ is the true mean value of the population of possible results (from which the observed results x_{ijk} are assumed to arise);

A_i is the error due to procedure/laboratory i;

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

ε_{ijk} is the random error for observation k in group ij.

The A_i are assumed to be drawn from a population with mean 0 and standard deviation σ_{L}, B_{ij} from a population with mean 0 and standard deviation σ_{bb} and ε_{ijk} from a population with mean zero and standard deviation σ_{r}. As before, the measurement error term is assumed to have the same variance, not only within each RM unit, but also within each laboratory. It is usually also assumed that the populations are normally distributed. The parameters to be obtained are the corresponding estimated between-laboratory standard deviation s_L, between-unit standard deviation s_{bb} and repeatability standard deviation s_r. Formula (B.7) shows how they are related to the error terms.

$$s_L = \sqrt{\mathrm{Var}(A_i)}$$

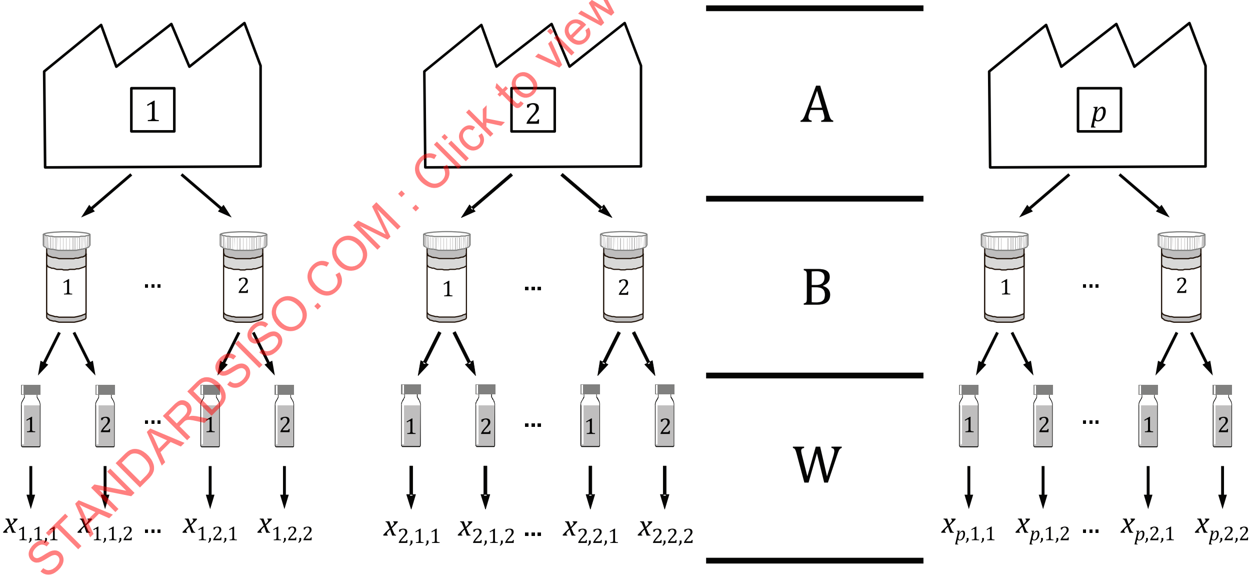
$$s_{bb} = \sqrt{\mathrm{Var}(B_{ij})} \tag{B.7}$$

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

For the between-unit homogeneity (s_{bb}), the same considerations apply with respect to the inability to detect batch heterogeneity as for the homogeneity study itself (see 7.8).

All these parameters can be estimated simultaneously by the analysis of variance (ANOVA) method[55], if there are sufficient results of equal replication (the same number of replicate determinations from each unit and the same number of units per procedure/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-unit 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 (see, for example, References [15], [57], [58] and [59]). A discussion of ANOVA in the context of the certification of reference materials is given in the literature (see, for example, References [16], [55], [60] and [61]).



Key

A between-laboratory variation

B between-unit variation (large containers indicate RM units)

W measurement repeatability (small containers indicate individual extracts)

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

The formulae for computing s_L, s_{bb}, and s_r are as follows[13],[16],[62]. The grand mean $\bar{\bar{x}}$ is computed using Formula (B.8):

$$\bar{\bar{x}} = \frac{1}{\sum_{i=1}^{p}\sum_{j=1}^{b_i} n_{ij}} \sum_{i=1}^{p}\sum_{j=1}^{b_i}\sum_{k=1}^{n_{ij}} x_{ijk} \tag{B.8}$$

where

p is the number of laboratories;

b_i is number of units used by procedure/laboratory i;

n_{ij} is the number of replicate measurements on unit ij.

s_L, s_{bb}, and s_r are given by

$$s_r^2 = M_{\text{within}} \tag{B.9}$$

$$s_{bb}^2 = \frac{M_B - M_{\text{within}}}{n_0} \tag{B.10}$$

$$s_L^2 = \frac{M_L - n'_0\, s_{bb}^2 - M_{\text{within}}}{(nb)_0} \tag{B.11}$$

where

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

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

and

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

The mean squares (M) should be obtained using published methods (for example, References [16] and [62]) or a suitable 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[14] may be used.

Following computation, s_{bb}, as calculated here, may be used for the assessment of uncertainty associated with heterogeneity in the same way as that calculated in Clause 7.

When applied in an interlaboratory context, it is important to check the assumption of constant within-group variance. This assumption can be checked using statistical tests such as Bartlett's or Levene's tests[57],[62].

NOTE The same model can be used to process data from a study of within-unit homogeneity of the form illustrated in Figure 5 b), by replacing references to laboratory with RM unit and references to RM unit above with references to within-unit aliquot. In a study of within-unit homogeneity, Formula (B.11) estimates the between-unit term s_{bb}, Formula (B.10) estimates the within-unit term s_{wb}, (which can be used as a component u_{wb} of the uncertainty for the certified value) and Formula (B.9) estimates the measurement repeatability standard deviation.

B.3 Linear regression (univariate linear model)

B.3.1 The basic model

Linear regression is used in reference material studies to determine simple rates of change and test their statistical significance. This clause describes the simplest case.

The simple linear model[56] can be expressed as in Formula (B.12):

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{(B.12)}$$

where β_0 and β_1 are the intercept and slope, respectively, and ε denotes the random error component, usually assumed to be normally distributed with a mean of zero.

Given a set of n pair-wise observations of Y versus X, individual observations (x_i, y_i) are related by Formula (B.13):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{(B.13)}$$

EXAMPLE For a basic stability study, x_i corresponds to an observed time and y_i to the corresponding observed value of the property of interest.

B.3.2 Fitting the model

The regression parameters can be computed from Formulae (B.14) and (B.15):

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{B.14}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{B.15}$$

where b_0 and b_1 are the estimated intercept and slope, respectively, and $\bar{x}$ and $\bar{y}$ are the mean of the respective observations.

The standard errors $s(b_1)$ and $s(b_0)$ in b_1 and b_0 can be computed using Formulae (B.16) to (B.18):

$$s(b_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{B.16}$$

where

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - b_o - b_1 x_i)^2}{n-2} \tag{B.17}$$

and

$$s(b_0) = s(b_1)\sqrt{\frac{\sum_{i=1}^{n} x^2}{n}} \tag{B.18}$$

B.3.3 Inspection and check of assumptions

Typical review of the preliminary fitted model includes:

- calculation of residuals [$r_i = y_i - (b_0 + b_1 x_i)$] and plotting against time, followed by inspection for evidence of curvature, run-to-run effects, substantial differences in dispersion from one time point to another, or for outlying values;
- preparation of a normality (or “Q-Q”) plot of residuals and inspection for evidence of non-normality;
- optionally, tests for significant differences in dispersion, significant between-group differences in residuals, or for nonlinearity.

B.3.4 Testing for statistically significant change

For the simple case above, after confirming the validity of the basic regression assumptions, the usual test for a statistically significant gradient is a *t*-test for slope significantly different from zero. This is conducted by calculating the *t* statistic in Formula (B.19):

$$t_{b_1} = \frac{|b_1|}{s(b_1)} \qquad \text{(B.19)}$$

and comparing this with the two-tailed critical value of Student's *t* for *n* − 2 degrees of freedom at the 95 % level of confidence. If the calculated test statistic t_{b_1} exceeds the critical value, the slope is considered to be significantly different from zero at the 95 % level of confidence.

B.3.5 Confidence interval for the regression line

The fitted values b_0 and b_1 may be used to provide an estimate $\hat{y}$ of the value of the response for a particular value $\hat{x}$, using

$$\hat{y} = b_0 + b_1\hat{x} \qquad \text{(B.20)}$$

It is sometimes useful to obtain a confidence interval for the predicted value $\hat{y}$. A two-sided confidence interval at the 95 % level of confidence is given by

$$\hat{y} \pm t_{95,n-2}\, s\sqrt{\frac{1}{n} + \frac{\left(\hat{x} - \bar{x}\right)^2}{\sum\limits_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} \qquad \text{(B.21)}$$

where $t_{95,n-2}$ is the two-tailed critical value for Student's *t* with *n*−2 degrees of freedom. Confidence intervals for different levels of confidence may be similarly obtained by using the appropriate critical value for Student's *t*.

NOTE This interval is not the prediction interval for individual values y_i.

B.4 Predicting shelf life or choosing initial monitoring point based on stability study results

B.4.1 Principle

Where there is an acceptable range, described by upper and/or lower specification limits, for change in a property value over time in storage, it can be useful to estimate either the storage lifetime (or 'shelf life') of a reference material, or set a first monitoring point, using the results from a stability study.

The principle is to estimate the 95 % confidence interval for future values, taking account of the estimated degradation rate, and to choose the shortest time at which one of the confidence limits (upper or lower) intersects a specification limit. This principle is illustrated schematically in Figure B.2.

For a CRM, any acceptable range chosen for the certified value should be less than the expanded uncertainty, ideally less than $U_{\text{CRM}}/3$. For other reference materials, any acceptable change should be sufficiently small to avoid adverse effects on the intended use.

B.4.2 Prediction of shelf life in the case of a linear trend

Let the upper and lower limits of an acceptable range of values for an RM property be L_{upr} and L_{lwr}, respectively. The confidence interval around a predicted future value of an RM when the degradation follows a simple linear form is given by Formulae (B.20) and (B.21), taking *x* as the time from the