International Standard

ISO 33406

First edition
2024-05

Approaches for the production of reference materials with qualitative properties

Approches pour la production de matériaux de référence avec des propriétés qualitatives

Reference number
ISO 33406:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

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

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 document 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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

For an explanation of 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 http://www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 334, *Reference materials*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

In 2015, ISO/TC 334 (formerly ISO/REMCO) published ISO/TR 79, which summarized the state of the art of the production of reference materials (RMs) with qualitative properties. The lack of an international vocabulary for terms and definitions for qualitative properties added an extra challenge in 2015 and ISO/TR 79 did not define or limit the use of the term "qualitative property". "Qualitative property" was therefore used as superordinate for nominal and ordinal properties. "Categorical properties" is a synonym for "qualitative properties". ISO/TR 79 lists examples of RMs that either have a certified value for a qualitative property or can be considered as in-house RMs characterized for a qualitative property. The examples listed are based on the principles elaborated in ISO 33405 and ISO Guide 80, but ISO/TR 79 did not undergo a consensus-building process.

For RMs, nominal properties are a particular kind of qualitative property (the other kind are ordinal properties). The identity of a polychlorinated biphenyl (PCB) and the species of a maple tree (genus *Acer*) are nominal properties where the values of the properties are the particular chemical species name or biological species name (e.g. PCB 105 and *Acer saccharinum*, respectively). The former has more than 200 possible values, the latter more than 160. The only meaningful comparison between values of a nominal property is whether they are identical or different.

Ordinal properties have values that can be ordered (i.e. ranked) from smallest to largest or from lowest to highest, but for which neither differences nor ratios are meaningful, even when their values are represented numerically.

EXAMPLE 1 Stage, as defined by the American Joint Committee on Cancer, is a property of solid tumour cancers (e.g. breast, colon or lung), whose possible values are the Roman numerals I, II, III and IV. However, one is not entitled to say either that stage IV is two times “worse” than stage II or that the difference in severity between stages III and I is the same as the difference in severity between stages IV and II.

EXAMPLE 2 The Mohs hardness of a mineral is expressed relative to a scale ranging from 1 (for talc) to 10 (for diamond) and can include half-integer values (e.g. 5 ½ for enstatite). However, fluorite (4) is not two times harder than gypsum (2), nor is the difference in hardness between topaz (8) and apatite (5) the same as the difference in hardness between quartz (7) and fluorite (4).

There are no unambiguous rules for expressing qualitative properties in various fields such as chemistry and biology. The following examples illustrate the confusion that exists. Qualitative properties can be described in different ways depending on the individual property or even on the way in which a specific property is expressed. For example, "colour" can be seen as a qualitative property of which "red" is a value. Alternatively, it has been proposed that "colour" is a general property of which "red" is an individual case. Moreover, a pure colour can be described by the corresponding wavelengths of light, with a band from 625 nm to 740 nm for red light. Hence, the qualitative property "colour" can be assigned a quantitative (or semi-quantitative) value such as 700 nm.

Qualitative value assignments can also differ according to the intended use. For example, "ethanol" may be treated as the identity of a specific compound or as an instance of a family of compounds with the general property "alcohols", which in turn can be seen as an instance of the general property "chemical species".

Seeing the progress and the increasing number of RMs characterized or certified for qualitative properties, ISO/REMCO decided in 2018 to start drafting internationally harmonized guidance for the production of such RMs.

This presents several issues with respect to terminology. The ISO Guide 30 definition of a certified reference material (CRM) requires it to have an RM certificate that provides the values of the specified properties, associated uncertainties and statements of metrological traceability. The relevance of metrological traceability to RMs with qualitative property values can, however, be unclear. Notes of the RM and CRM definitions in ISO Guide 30 clarify that the concept of value includes a qualitative property, such as identity or sequence, and that uncertainties for qualitative values can be expressed as probabilities or levels of confidence[18]. However, a qualitative property has no numerical value and at present ISO/IEC Guide 98-3 (GUM) provides no methodology for assigning an uncertainty to such property values. Metrological traceability is defined in ISO/IEC Guide 99 (VIM) as a property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations. Hence, it too is inapplicable to qualitative property values. Nevertheless, RMs with qualitative properties

are needed now and are being produced in increasing numbers. In order to provide meaningful guidance on these aspects of RM production, this document:

— describes how confidence in qualitative assigned values can be expressed;

— describes the relevance of a documented unbroken chain of qualitative comparisons in assigning qualitative values by comparison with qualitative references or reference information;

— notes that determination of many qualitative properties depends on data obtained with quantitative measurements and that measurement uncertainty and metrological traceability are relevant to these measurements.

These issues are discussed in detail in 5.2.2 and 5.3.2.

General requirements, structural requirements, resource requirements and management requirements for the production of RMs are described in ISO 17034. These requirements apply to the production of RMs with qualitative properties. This document supplements ISO 17034 and related guidance on the production of RMs by providing additional guidance on the value assignment and the assessment of homogeneity, stability and commutability for RMs with qualitative properties.

Approaches for the production of reference materials with qualitative properties

1 Scope

1.1 This document notes the requirements of ISO 17034 and provides guidance on the implementation of ISO 17034 in the production of RMs having one or more assigned qualitative property values, for expressing uncertainties for qualitative property values, and for establishing traceability.

NOTE The concepts of traceability and uncertainty address characteristics similar to those addressed by the concepts of traceability and measurement uncertainty as used in the metrology of quantitative properties.

1.2 This document therefore does not describe aspects related to the production of RMs with quantitative property values.

NOTE Annex A provides examples of types of RMs within the scope of this document.

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 17034, *General requirements for the competence of reference material producers*

ISO 33401, *Reference materials — Contents of certificates, labels and accompanying documentation*

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)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 17034, ISO 33401, ISO Guide 30, and ISO/IEC Guide 99 apply.

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

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

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

4 Qualitative properties

Qualitative properties, also called categorical properties, can be nominal or ordinal. A nominal property has values that divide the set of materials that have it into classes such that all the materials in the same class have the same value of the property, and the only comparison that can be made between values of the property is of whether they are identical or different. An ordinal property is similar, except that the comparisons that can be made between two values of the property are of relative rank order, i.e. whether one is lower, equal to or higher than the other. Qualitative properties can have two or more possible values. This document focuses on nominal properties.

5 Meeting technical and production requirements

5.1 Characterization

5.1.1 General considerations

5.1.1.1 As outlined in ISO 33405 and ISO/TR 79, materials can be characterized for qualitative properties such as colour, odour or shape. In some cases, the result of a qualitative characterization can be expressed as a qualitative or quantitative property value. Examples are particle shapes or colour according to the Hunter system[19]. This transforms the problem to the characterization of a quantity, which could be an operationally defined measurand. For colour especially, characterization of the absorbance or reflectance spectrum may also be considered.

5.1.1.2 Many qualitative properties are not meaningfully quantifiable, for example identity of a substance, species of an organism or gender of an animal.

5.1.1.3 Approaches to characterization for qualitative properties include:

— characterization using one or more qualitative determinations;

— characterization based on provenance (5.1.2);

— characterization using measurements of quantitative properties (5.1.3);

— characterization by a combination of methods (5.1.4).

5.1.2 Materials characterized based on provenance

5.1.2.1 An RM 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 reference material producer (RMP) should obtain documentary or other evidence of the origin of the material that shows an unbroken chain of evidence from origin to final packaging. All evidence should be retained for the lifetime of the material.

NOTE The term 'provenance' is used here in the sense of origin or place of origin. The term can also apply to the evidence of the origin or place of origin.

5.1.2.2 RMPs should have procedures in place to ensure that the provenance is maintained. These procedures should include handling of the material (e.g. sampling, homogenization, packaging, storage) and prevention of contamination by other materials.

NOTE For biological materials, provenance can include evidence of parentage or continuous culture from an authentic specimen (including an RM).

5.1.2.3 Characterization based on provenance should be supported by additional evidence to confirm identity of the material (see, for example 5.1.3, 5.1.4).

5.1.3 Materials characterized for identity based on measurements

5.1.3.1 General consideration for materials characterized for identity based on measurements

As outlined in ISO 33405, when characterizing the identity of a substance based on measurements, several aspects should be considered, including the following:

— Characterization can be based on measurement results from one or several methods. For example, chemical shifts and areas of peaks in a nuclear magnetic resonance (NMR) spectrum or a combination of colour, melting point, relative molecular mass, etc.

— 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.

— 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.

— 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 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, histopathological examination). However, this judgement is usually based on observations and comparison with characterization criteria and, for example, RMs or reference data recognized by the relevant community of users. Expert judgement based on observations falls within the scope of 5.1.3.1; it can be used, for example, in identification of the species of a plant based on comparisons with voucher specimens.

5.1.3.2 Criteria for characterization of identity by measurement

5.1.3.2.1 As outlined in ISO 33405, testing for identity of a material involves comparison of a set of measurement results on that material with predefined acceptance criteria (e.g. melting point range, degree of similarity with a reference DNA sequence) for these measurement results.

NOTE Sequence similarity can be reported as "percent identity", i.e. the percentage of characters that match between two sequences.

EXAMPLE An organic polymer material could 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 % relative absorbance;

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

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

5.1.3.2.2 Sources of criteria can include internationally recognized compendia (e.g. pharmacopoeia sources[20] and other collections of reference data[21]). Such information can change outside the control of the RMP. RMPs should therefore clearly state the criteria used for the assignment of identity, either as a set of values or as a dated reference on the RM document to applied criteria.

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

5.1.3.2.4 An assignment of identity (of a substance, biological species, etc.) should be certified only when such assignment is made beyond sufficiently high confidence for the intended use (see 2.11 in Reference [22]).

5.1.3.2.5 Comparison of results with predefined acceptance criteria also applies to other qualitative properties characterized using measurement.

5.1.4 Characterization by a combination of methods

5.1.4.1 As outlined in ISO 33405, 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 RM. Frequently used methods include, for example, determination of melting point, relative

molecular mass, ultraviolet (UV), IR, NMR and mass spectra. Sensory methods may also be applied. Together with information on the raw material and its processing steps and the sampling and transport to the RMP, the collection of methods should be sufficient to establish the identity of the material. If detailed published criteria (e.g. pharmacopoeial criteria for identification) exist, the choice of methods can be restricted to those listed.

NOTE 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.

EXAMPLE 1 When using a standard strain of a bacteria from a recognized culture preservation centre, macroscopic features (characteristics on typical media), microscopic features (e.g. characteristics of gram stain, identification of coccus or bacillus), phenotypic features (characteristics in biochemical reactions) and other features where applicable (e.g. plasma coagulase assay, antioxidant enzyme test, serology test) can be necessary in the documentation.

EXAMPLE 2 A seed of a plant obtained from an institute with only morphological features will possibly not support the taxon of this plant; it can require further evidence or documentation.

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

5.1.4.3 The results of each of the tests and measurements made should be compared with the criteria for the proposed substance. Published procedures for 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 published criteria, identity is established with a negligible uncertainty.

NOTE A judgement on whether the accumulated measurement and provenance information is sufficient to establish identity is somewhat subjective. Peer review can help to increase confidence in the assignment.

5.1.4.4 Where applicable, a suitable assessment or statement of purity of the RM should be made.

NOTE When a combination of qualitative and quantitative determinations is used for characterization, it can be useful to use similarity measures to support characterization. ISO/TR 79 gives an example for similarity measures.

5.2 Application of metrological traceability to qualitative determinations

5.2.1 General

Metrological traceability is defined in ISO/IEC Guide 99 as a "property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty." This approach for achieving comparable results is applicable to quantitative measurements, including where quantitative measurement results are used to determine qualitative characteristics of materials. It cannot be applied directly to qualitative property values in accordance with the terminology in ISO/IEC Guide 99 since these are not measured values based on calibration. Nevertheless, the need for consistent measurement data remains and the guidance in 5.2.2 to 5.2.4 describes several ways in which this can be achieved. First, many assignments of qualitative property values depend on measurement of one or more quantitative values for which metrological traceability can be stated in the usual way. Second, traceability of qualitative property values may be established by comparison with a reference. For qualitative properties, the reference to which traceability is claimed will not be a unit of the International System of Units (SI) but some reference recognized by the relevant scientific community, including users of the RM. This can be an artefact such as a CRM or can take the form of authoritative data such as, for example, published photographs and associated descriptions. Finally, qualitative properties can be assigned through the provenance of the material. For example, if an RM is claimed to be characteristic of a specific plant, its provenance should be sufficient to confirm its origin. Documentation of the provenance in this way is sometimes referred to as trackability but it fulfils the same aims for qualitative values as metrological traceability for quantitative values. These three approaches are described in more detail in 5.2.2.

5.2.2 Metrological traceability

5.2.2.1 Metrological traceability applies to the determination of quantitative properties, including conditions for tests yielding a qualitative result. Many qualitative value assignments include the measurement of one or more quantitative characteristics, such as melting point or pH. Others can include the application of test procedures which include quantitative conditions, such as times, temperatures and concentrations. The need for statements of metrological traceability for these parameters is discussed in 5.2.2.2 and 5.2.2.3.

5.2.2.2 Where a qualitative value assignment depends upon measurements or tests that include quantitative conditions, the measurements and/or tests should meet the metrological traceability requirements of ISO/IEC 17025.

EXAMPLE 1 In a structure assignment using spectroscopic methods, the wavelength or frequency scale describing the location of signals is checked or calibrated.

EXAMPLE 2 In a qualitative analytical test which requires application of a reagent at a stated concentration, the reagent is prepared using appropriately calibrated equipment to provide a reliable concentration.

5.2.2.3 Where a qualitative value assignment depends upon measurements or tests that include quantitative conditions, a statement of metrological traceability relating to the relevant measurements and/or tests should be included on the RM certificate.

NOTE ISO/TR 16476 and Reference [23] provide further information on the expression of metrological traceability.

5.2.3 Reference data and reference materials for qualitative determinations

5.2.3.1 Value assignment for qualitative properties can depend on comparison with reference data or with another RM already value-assigned for the qualitative property concerned. An appropriate RM for value assignment can be, for example, a sample whose identity has been reliably established or an independently characterised example of the class of interest.

EXAMPLE A herbal material is identified based on microscopic comparison with a reference sample held by a national repository or archive.

5.2.3.2 Where value assignment of a qualitative property depends critically upon comparison with particular reference data or a particular RM, the reference should be stated in the RM documentation accompanying the RM being produced.

NOTE Value assignment depends critically on a particular comparison when the particular comparison is essential to the assignment of the qualitative value.

5.2.4 Qualitative value assigned based on provenance

5.2.4.1 Value assignment based on provenance – knowledge of the origin of the RM – is described in 5.1.2. In the case of value assignment based on provenance:

— the RMP should maintain full documentation of the provenance of the material, including the source of the material, changes in ownership and procedures to avoid contamination or replacement of the material with any other;

— the basis of the assignment, including the source of the material, should be made available in the RM document.

5.2.4.2 Where assignment is based upon provenance supported by confirmatory measurements or tests, the provisions of 5.2.2 and 5.2.3 apply for the confirmatory measurements or tests.

5.3 Measurement uncertainty and confidence in qualitative values

5.3.1 General considerations

5.3.1.1 Every assignment of value to a property is surrounded by uncertainty. The techniques for uncertainty evaluation described in ISO/IEC Guide 98-3 are not applicable to evaluate uncertainties associated with value assignments to qualitative properties. Nevertheless, it is important to provide users of these RMs with guidance on the reliability of the assigned value. Many assignments of qualitative (nominal) properties depend on measurement of one or more quantitative values for which usual measurement uncertainty estimates are available. There is also a wide range of other information that can allow the user to assess confidence in the assigned value. It is important to document all such information.

5.3.1.2 While measurement uncertainty is well defined (see ISO/IEC Guide 98-3) and applies unambiguously to the values of quantities, the meaning of “uncertainty” as it applies to values on ordinal scales or to nominal values is not well established. There is also little or no harmonized guidance on conveying the degree of uncertainty, or the degree of confidence, that the user can have in the assigned value of a qualitative property. Pending a harmonised framework, the principles adopted in this document are as follows:

— The term "confidence" refers to the degree of belief one has in the value assigned to a qualitative property. Such confidence may be expressed qualitatively, using an ordinal scale (e.g. "most confident", "very confident", "moderately confident") or quantitatively (e.g. as a likelihood ratio or as a probability distribution on the set of possible values of the qualitative property).

— Users of RMs with qualitative values should have sufficient confidence in the values provided for the intended use of the material. In particular, where assigned qualitative values are certified the RMP should clearly state the justification for their confidence in the value.

— All assignments of value to qualitative properties should be qualified with a statement of confidence, even if this statement is itself qualitative and expresses a subjective expert opinion. Quantitative statements of confidence in assigned qualitative values are not required but are permitted where they do not give a misleading impression of reliability of the value.

NOTE Annex B gives further information on expression of confidence for qualitative values.

5.3.2 Measurement uncertainty

5.3.2.1 Measurement uncertainty as described in ISO/IEC Guide 98-3 impacts qualitative measurements in two ways:

— Control of uncertainties in test conditions, such as times, temperatures or lengths, is important for reliable qualitative value assignment when it involves measurements of quantities or control of (quantitative) test conditions.

— Measurement uncertainty related to intermediate measurements can contribute to the estimation of false response rates; for example, where classification depends on measurement results exceeding a threshold.

5.3.2.2 Where a qualitative value assignment depends upon measurements or upon tests that include quantitative conditions, the RMP should ensure that the measurement uncertainties are sufficiently small that they have no significant impact on the confidence in the assigned qualitative value. This should include one or more of the following:

— Control of conditions affecting the test result to within well-established and documented tolerances.

— Demonstration that the uncertainty is sufficiently small to have no significant influence on the outcome of the test.

NOTE Because low false response rates are hard to determine, it is rarely practical to do more than show that substantial change in a condition – for example, larger than three times the uncertainty – has limited or no detectable effect on false response rate.

5.3.2.3 Where assignment of a qualitative value depends upon comparison of a measured value with one or more limits, the RMP may use the measurement uncertainty, together with suitable assumptions about the distribution of results, to estimate the probability of false classification.

5.3.2.4 RMPs are not required to report measurement uncertainty information for intermediate measurement results on a RM document that relates solely to assignment of qualitative values.

5.3.2.5 The quantitative expression of uncertainty may be a probability distribution for the property of interest or it may be a summary characterization of how dispersed this distribution is. Alternatively, the uncertainty may be expressed using measures of confidence derived from false response rates (false positive, false negative) such as sensitivity and specificity (see Table B.2). Examples are given in B.4.

5.3.3 Confidence in qualitative values

5.3.3.1 RMPs should provide sufficient information with the RM to allow the user to assess whether the confidence claimed for the assigned value meets their needs.

5.3.3.2 RMPs are not required to provide quantitative statements of confidence in assigned qualitative values in the RM document for the material but may do so in order to provide sufficient information with the RM to assist the end user to judge whether the RM is suitable for the end user's needs.

NOTE A quantitative statement of confidence in an assigned qualitative value can be a probability or level of confidence.

5.3.3.3 Information that assists the user to assess the claimed confidence in the assigned value may include:

a) the basis for the assigned value (e.g. provenance, testing, expert opinion);

b) a description of the origin of the material;

c) a description of procedures for minimising contamination or inadvertent interchange of materials;

d) information on the nature and performance of qualitative test methods used in value assignment;

e) information on reference data used to support value assignment;

f) a qualitative or verbal indication of confidence;

g) a quantitative indication of confidence in the assigned value.

NOTE 1 Subclause B.2 gives further information on performance figures for qualitative tests.

NOTE 2 Subclause B.4 gives further information on quantitative indications of confidence in an assigned qualitative value.

5.3.3.4 Where test performance information is provided (see 5.3.3.3, d), the performance information should be presented in a manner that does not give a misleading impression of confidence in the assigned value.

EXAMPLE An identification using nucleic acid sequencing provides a theoretical chance match probability below 10^{-50}. In practice, the probability could be considerably higher. This can be useful information for a knowledgeable practitioner. However, the chance of errors during packaging or certificate preparation cannot readily be reduced to such low levels. It is then unsafe for the user to infer that the probability of incorrect value assignment is as low as 10^{-50}. Inexperienced users of the material can then require cautionary advice on the relevance or interpretation of the test performance data, for example, using a qualitative indication of the strength of evidence using an ordinal scale (see Annex B.1).

5.3.3.5 A qualitative indication of confidence (see 5.3.3.3, f) may be, for example, a brief explanatory indication on a rating scale or an abbreviated statement indicating a basis for assignment. Where a value on a rating scale is included as an indication of confidence on a certificate, the complete scale should be available to the user of the material. The qualitative evaluation should be based on explicit criteria that are specifically appropriate and unambiguous for application in the relevant context and which should be stated in the corresponding certificate.

EXAMPLE A CRM comprises leaves of *Ginkgo biloba*. The uncertainty associated with the identification of the biological species was expressed as a verbal statement on a rating scale (see B.1). A qualitative value based on a qualitative test result is accompanied by a statement of the form:

> "The test result provides very strong support for the assigned value, using the scale weak, moderate, strong, very strong."

NOTE Sometimes such scales are derived from numerical values of, for example, a likelihood ratio or a probability (B.4).

5.3.3.6 A quantitative expression of uncertainty may be a probability distribution over the set of possible values for the property of interest or it may be a summary characterization of how dispersed this distribution is on the set of those possible values. Alternatively, the uncertainty may be expressed using classification error rates (false positive, false negative), sensitivity and specificity (see Table B.2) or receiver operating characteristics.

EXAMPLE Some DNA sequencing techniques characterize the confidence in the identification of the nucleobase occupying a particular locus in terms of integer quality scores, one for each of the four possible identities – A, C, G or T – that the nucleobase has.

NOTE Subclause B.4 gives further information on the assessment of quantitative indications of confidence in an assigned qualitative value.

5.4 General considerations for the selection of statistical approaches

5.4.1 In contrast to the well-established range of statistical approaches for value assignment, homogeneity testing, stability testing and commutability assessment for quantities, few statistical approaches have been widely applied to the characterisation of qualitative properties. 5.4 to 5.6 accordingly provide general guidance on statistical approaches that have been found useful, or proposed, for particular tasks. The list is not exhaustive and other approaches may be applied for particular circumstances.

5.4.2 In selecting statistical approaches, the following considerations are particularly important in applications for RMs with qualitative properties:

— Qualitative tests generally provide less information per test result than measurements on quantities that return a numerical result. Wherever possible, therefore, experimental methods that return quantitative information related to the qualitative properties of interest should be preferred to tests that return qualitative information. For example, it can be possible to conduct a stability assessment using the concentration of a particular protein that indicates a biological species, rather than relying on a qualitative test procedure that tests for the presence of the protein.

— When a qualitative test method has an appreciable false response rate, it can be particularly difficult to distinguish qualitative results that arise from failures in the testing process from qualitative results that arise from defects or changes in a candidate RM. Test methods for characterization, homogeneity assessment and stability assessment should therefore be chosen for high reliability when possible.

— RMs are intended to be sufficiently homogeneous and stable. Discordant results (i.e. results which are not as expected for the property under consideration) are expected to be rare and can be completely absent. For example, with reliable testing methods, a stability check on a particular characteristic could return a positive result for all items and replicates, with no negative results. Statistical methods for analysis of qualitative test results should accordingly be chosen to accommodate low or zero proportions of discordant responses.

5.5 Assessment of homogeneity

5.5.1 General considerations

5.5.1.1 According to ISO 17034, the RMP must assess the homogeneity of the RM in its final packaged form to ensure its fitness for purpose. The RMP needs to address between-unit homogeneity for batch certification and within-unit homogeneity. The latter is usually addressed by the minimum sample size.

5.5.1.2 The homogeneity of the material should be assessed for all the assigned property values.

5.5.1.3 Homogeneity of the assigned qualitative property value(s) in an RM should be assessed as response to test methods directly targeting the assigned nominal property value(s) and to any other methods listed in the intended use section of the RM document.

NOTE Homogeneity of the properties that are measured to support the qualitative value assignment by the RMP does not necessarily guarantee homogeneity of all properties that the end user can measure using the RM, even if the test method belongs to the same category of methods, since each property can have a different degree of homogeneous distribution.

5.5.1.4 When the assignment of qualitative property value(s) is based upon provenance, supported by confirmatory measurements or tests of further properties, homogeneity of the RM should be assessed by the RMP following ISO 33405, and methods targeting these properties of the RM should be selected.

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

5.5.2 Experimental designs for homogeneity assessment for qualitative values

5.5.2.1 ISO 33405 provides general experimental designs for homogeneity studies, including sampling strategies for obtaining a representative selection of RM units. These designs and sampling strategies may be applied to the assessment of homogeneity for qualitative properties. For experimental studies of homogeneity using a quantitative property, the provisions of ISO 33405 for experiment design apply.

NOTE Further guidance can be found in Reference [18].

5.5.2.2 The following additional considerations apply when the homogeneity study uses a test method that returns a qualitative result:

a) The number of RM units studied should be increased to provide a sufficiently narrow confidence interval for the estimated proportion of defective units (see 5.5.2.3) or, where using sampling plans for inspection by attributes, should be selected according to the chosen sampling plan (see 5.5.3.3).

b) Where the chosen test method is known to have an essentially negligible false response rate (so that all compliant RM units will return a compliant result for any single test on a compliant RM unit, or a non-compliant result for any non-compliant unit), replicate testing may be omitted. That is, the experiment may consist of one test on each chosen RM unit.

c) The minimum number of units to be sampled for confidence probability analysis depends on the type of criteria for providing the qualitative test results[18].

NOTE In case b), between-unit inhomogeneity cannot be distinguished from within-unit inhomogeneity or from chance failures in the test method, unless additional information on, for example, test method variability is available. Absence of differences between units can, however, be interpreted as absence of evidence of inhomogeneity.

5.5.2.3 As outlined in ISO 33405 , for RMs certified for qualitative properties, 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 or similar guidance.

NOTE For materials with assigned qualitative properties, an experimental homogeneity study is likely to be limited to a check for unexpected gross heterogeneity (e.g. greater than 10 % defective), with other information on, for example, origin of material or processing used to support any statement of homogeneity. 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 RM batch sizes.

5.5.3 Statistical approaches for homogeneity assessment for qualitative properties

5.5.3.1 Principles and procedures

5.5.3.1.1 Two general approaches can be applied:

a) Confirmation that the proportion of defective RM units does not exceed a chosen threshold. This can be applied when replicate tests are performed on each RM unit or when only one test is applied to each RM unit.

b) Testing for a significant between-unit difference in the probability of non-compliant responses. This applies when replicate tests are performed on each RM unit.

5.5.3.1.2 Approach a) in 5.5.3.1.1 can be implemented by calculation of an upper confidence limit for the proportion of non-compliant RM units. This is described in 5.5.3.2. Alternatively, the principles of inspection by attributes can be applied; this is discussed in 5.5.3.3.

5.5.3.1.3 Approach b) in 5.5.3.1.1 can be implemented by several statistical procedures. Two suitable procedures are summarized in 5.5.3.4 and 5.5.3.5.

5.5.3.2 Confidence interval for proportion of non-compliant RM units

5.5.3.2.1 The RMP should set a criterion for a proportion of defective RM units that would require the RM to be discarded or further processed.

5.5.3.2.2 For any of the designs in 5.5.2, a one-sided upper confidence limit for proportion of non-compliant RM units should be calculated. This provides summary information for homogeneity assessment. Confidence interval procedures for low proportions should be used. Appropriate procedures include the "exact" Pearson-Clopper binomial interval[24] and the Wilson scoring interval[25]. Where tests are replicated for each RM unit, the interval should be based on the number of RM units considered defective and not on the total number of responses. In the special case where no failures are observed, a simple approximate interval for proportion of defective RM units is $[0, 3/n]$, where n is the number for RM units tested.

5.5.3.2.3 The calculated confidence limit should be compared with the criterion set in 5.5.3.2.1. If the confidence limit is less than or equal to the chosen criterion, the material may be accepted as sufficiently homogeneous.

NOTE 1 This procedure is a test for positive compliance with a criterion. It provides assurance that the proportion of defective RM units is below the level at which reprocessing would be considered necessary.

NOTE 2 Subclause C.2 describes the calculation of the Wilson interval.

NOTE 3 Subclause D.1.4 provides an example of the calculation of the Wilson interval for a homogeneity study.

5.5.3.3 Inspection by attributes

5.5.3.3.1 Sampling and acceptance schemes for inspection by attributes are described in ISO 2859-1 and ISO 2859-4. They are intended for repeated batch (or “lot”) inspection and can be particularly useful when RMs are produced in relatively frequent repeated batches. The basic principles are:

a) Set an "acceptance quality limit" (AQL) which is a "worst tolerable process average (for the proportion of defective units) when a continuing series of lots is submitted for acceptance".

NOTE The AQL is a parameter of the sampling scheme and not to be confused with the process average that describes the operating level of the RM production process. It is expected that the process average will be better than the AQL to avoid excessive rejections.

b) Choose an "inspection level". An inspection level is a level of stringency for the inspection procedure. ISO 2859-1 and ISO 2859-4 recommend the "normal" level or declared quality level 0 (zero) for most purposes.

NOTE In repeated lot inspection, ISO 2859-1 adjusts the inspection level depending on the number of recent lots that pass or fail inspection.

c) Choose a sampling plan based on the number of RM units in the production batch (the "lot size" in ISO 2859-1). For a single production batch, a single sampling plan is appropriate.

d) Take a representative sample of the required number of RM units and test all the RM units in the sample.

e) If there are no failures (non-compliant RM units), the lot is accepted. If any failures are detected, the number of failures is compared with the acceptance and rejection numbers in the sampling plan. If the number of failures is greater than or equal to the rejection number, the lot is rejected.

NOTE 1 This procedure rejects a batch if the proportion of defective RM units is significantly greater than the AQL at approximately 95 % level of confidence. The actual proportion of defective RM units can therefore be considerably higher than the AQL.

NOTE 2 The upper 95 % confidence limit for the proportion of nonconforming units can be calculated using the procedure in C.2.

NOTE 3 Additional guidance on samples sizes for RMs with a lot size less than 250 can be found for example in GB/T 2828.11 and Reference [26].

EXAMPLE A batch of approximately 2 000 RM units is produced. No prior information on homogeneity exists. The acceptance quality limit is set at 1,0 (i.e. 1,0 % nonconforming items). The inspection level is set to normal (level II in ISO 2859-1). Reference to the single acceptance sampling plans in ISO 2859-1 shows that the batch size falls into “sample size code” K (1 201 – 3 200 items). The single sampling plan for normal inspection requires examination of 125 RM units. The acceptance number for AQL of 1,0 is then three (see ISO 2859-1) and the rejection number is four. A batch is therefore accepted if there are three or fewer non-compliant units and rejected if there are four or more.

5.5.3.3.2 In repeated batch production and inspection using single sampling plans, the inspection level should be increased to "tightened" if two of any five successive batches fail and may be reduced back to normal when five consecutive batches have passed under tightened inspection. ISO 2859-1 also provides for reduced inspection based on a cumulative score.

NOTE ISO 2859-1 provides detailed rules for changing inspection level.

5.5.3.3.3 If the false response rate is well known or other information of the target property value is available, the number of samples may be significantly decreased using the methods in ISO 2859-4.

5.5.3.4 Two-way contingency table analysis

5.5.3.4.1 For replicated tests on each of a number of RM units, a two-way contingency table is a table of counts of positive and negative results by RM unit. Statistical tests such as a chi-squared test or (for small

numbers of negative responses) Fisher's exact test[27] can be applied to test for significant differences in the probability of a negative response for different units.

5.5.3.4.2 For homogeneity testing, the RM should be regarded as inhomogeneous if the test statistic exceeds the 95 % critical value or, equivalently, if the corresponding *p*-value is below 0,05.

NOTE 1 Subclause C.1 provides a procedure for a chi-squared test applicable to a two-way contingency table.

NOTE 2 Subclause D.1.3 provides an example of the analysis of a homogeneity study using a chi-squared test.

5.5.3.5 Categorical analysis of variance (CATANOVA)

CATANOVA[28] or, for ordinal responses, ORDANOVA[29], can be used to test for a significant between-unit difference in probability of discordant results. Both procedures are applicable to experimental designs involving replicate tests on a number of RM units. Detailed calculation procedures are available for the general case[28] and in an application for homogeneity testing[30].

NOTE Further information on two-way CATANOVA and two-way ORDANOVA is available; see References [31,32].

5.6 Assessment of stability

5.6.1 General considerations for stability assessment

5.6.1.1 According to ISO 17034, the RMP must assess the stability of the RM with respect to the assigned qualitative value(s) given on the RM document. The RMP needs to assess the short-term stability to guarantee that suitable transport conditions are selected, the long-term stability to ensure appropriate storage conditions and the stability during the use by the end user. It is the responsibility of the RMP to assess all three conditions of stability and provide a period of validity for the qualitative value as well as instructions for use to the end user.

5.6.1.2 Stability of the assigned qualitative value(s) in a reference material should be assessed in terms of stability of response to test methods targeting the assigned qualitative property value(s).

5.6.1.3 Where experimental stability study is required, the RMP should use one or more of the following:

a) Determination of the change in proportion of units responding correctly to qualitative test methods for the qualitative properties of interest.

b) Study of quantitative properties that can indicate degradation of the material.

NOTE 1 It can be helpful for users of the material to provide information regarding the methods and properties used to investigate the stability of the material.

Note 2 Subclause 5.6.2 gives further details of appropriate designs for stability studies using both qualitative test methods (as in item a) in 5.6.1.3) or quantitative methods (as in item b) in 5.6.1.3).

5.6.1.4 When the assignment of qualitative property value(s) is based upon provenance supported by confirmatory measurements or tests, stability of the RM should be assessed by the RMP, following ISO 33405 , and methods targeting these properties of the RM should be selected.

NOTE It can be helpful for users of the material to provide information regarding the methods and properties used to investigate the stability of the material.

5.6.1.5 The RMP should provide an informed recommendation on the stability of the assigned properties of the RM during the period of validity. The RMP should give instructions for handling and use of the material on the RM document.

NOTE Stability of the properties that are measured to support the qualitative value assignment by the RMP does not necessarily guarantee stability of all properties that the end user can measure using the RM, even if the test method belongs to the same category of methods, since the material can have a different stability with respect to each property.

5.6.2 Designs for experimental stability studies for qualitative properties

5.6.2.1 Basic designs

5.6.2.1.1 ISO 33405 provides general experimental designs for stability studies, including sampling strategies for obtaining a representative selection of RM units. These may be applied to the assessment of stability for qualitative properties. For experimental studies of stability using a quantitative property, the provisions of ISO 33405 for experiment design apply.

5.6.2.1.2 The simplest experimental design for assessing the stability of quantitative or qualitative properties is to expose a large number of units to planned storage (or other) conditions and to test a fixed number of units at intervals. ISO 33405 refers to this as a classical stability study. A closely related alternative is the isochronous design, in which testing is deferred until the end of the study, after moving RM units periodically to reference conditions. In the case of qualitative properties, both the classical and isochronous design provide an estimated proportion of failed units at each of a series of times.

5.6.2.1.3 Additional factors, such as temperature and humidity, may also be varied systematically to give accelerated studies or to test the effect of additional factors.

5.6.2.1.4 The number of units chosen for stability studies depends on the design, the possible risk for the end user, the expected change in failure rate within the study time and the confidence required in the result. As a guide, when there is no prior evidence of stability available, simple classical and isochronous stability studies should include 10 RM units for each time point or 40 RM units for the complete study.

NOTE Ten observations per group is the smallest number at which a change from zero probability of failure to a 50 % probability of failure can be reliably detected using a two-sided, two-sample proportion test at a level of confidence of 95 %. This accordingly provides good assurance of detecting a large change with the smallest practicable number of time points.

5.6.2.2 Continuous monitoring design

An alternative design for experimental stability assessment approach is applicable where it is possible to monitor the condition of all units regularly and non-destructively. The approach is to reserve a large set of RM units, monitor all RM units continuously, and record the time at which each RM unit fails, up to a predetermined maximum time. This approach provides a set of known failure times, together with a number of RM units whose failure time is known to be greater than the duration of the study. This type of data can be assessed using methods from survival analysis. A brief description is given in 5.6.2.5.

5.6.2.3 Contingency table analysis

5.6.2.3.1 In the context of stability tests, the simplest contingency table is a table of the numbers of unaffected and failed RM units at each time point in a stability design (see 5.6.2.1), forming a two-way table. Contingency table analysis (using chi-squared or other hypothesis tests) then provides a statistical test for differences in proportion of failures at different times; a significant difference indicates possible instability. Additional factors, such as temperature, can be accommodated by additional cross-tabulation.

5.6.2.3.2 Care should be taken in interpreting the results of contingency table analysis for stability data. Contingency table analysis simply tests for differences between groups and does not use the ordering information or interval size for different time or temperature points. A significant result from a hypothesis test therefore does not necessarily indicate a consistent increase in failures with time. A significant result from a contingency table test should accordingly be followed up by inspection for consistent trends, for example using a plot of proportion failing over time.

NOTE 1 Subclause C.1 provides a procedure for a chi-squared test applicable to a two-way contingency table.

NOTE 2 Subclause D.2 provides an example of the use of contingency table analysis applied to a stability study.

5.6.2.4 Regression of failure probability against storage or exposure time

5.6.2.4.1 Logistic regression may be applied to counts or proportions of failed units or to individual failures. Individual failures are typically coded as 0 for unaffected RM units and 1 for failed or defective units. The simplest logistic regression model describes the change in probability of failure with time according to Formula (1).

$$ln\left(\frac{p}{p-1}\right)=b_0+b_1t \tag{1}$$

where

p is the probability of failure;

ln denotes the natural logarithm;

t is the elapsed time at particular storage conditions;

b_0 and b_1 are model coefficients.

This model is similar to the simplest linear model for stability assessment of quantitative properties.

5.6.2.4.2 Probit regression may be used as an alternative to logistic regression. Probit regression uses the model in Formula (2).

$$p=\Phi\left(b_0+b_1t\ldots\right) \tag{2}$$

where

p is the probability of failure;

Φ is the cumulative normal distribution function;

t is the elapsed time at particular storage conditions;

b_0 and b_1 are model coefficients.

5.6.2.4.3 In practice, logistic and probit regression are similar in form. The choice between one and the other may be made by considering:

— the underlying process leading to failure;

— the adequacy of the fit of the model to the data; and

— the availability of software.

NOTE 1 Adequacy of the fit of the model to the data can be evaluated using Akaike's Information Criterion (AIC) or Bayesian information criterion (BIC)[33].

NOTE 2 Logistic regression can be generalized to nominal properties taking more than two values, and to ordinal properties (as described in Chapter 8 in Reference [27]).

NOTE 3 Logistic regression and probit regression are kinds of generalized linear models. References [27,34,35] provide detailed descriptions of these models and of the computational methods used to fit them to data. Reference [36] provides an example of application.

NOTE 4 Logistic and probit regression will not generally be successful where a data set contains very few or no failures (or where almost all units have failed). These procedures are accordingly more useful in analysis of accelerated studies, in which some RM units are exposed to conditions that increase the probability of early failure.

5.6.2.5 Survival analysis

5.6.2.5.1 The statistical analysis of cumulative failures over time is known as survival analysis, from its historical use in clinical studies. It is sometimes also called reliability analysis, especially in engineering. Some of the methods are particularly appropriate to the analysis of data from the continuous monitoring design described at 5.6.2.2. Survival analysis is a wide field and it is not possible to describe individual methods in detail in this document. A brief summary of some of the principal models is given in Table 1, with some advantages and disadvantages. The methods can be grouped into parametric methods, which assume a particular model for the probability of failure over time, and nonparametric methods, which do not assume a particular model.

5.6.2.5.2 All survival analysis models accommodate "right-censored" data, i.e. failure times that are only known to be greater than the study duration. All require software for implementation.

Table 1 — Summary of selected survival analysis models

Model	Formula	Advantages and disadvantages
Logistic model	see Formula (1)	Simple model, frequently provided by statistical software. This is applicable to individual observations or proportions calculated for groups.
Exponential model[a]	$S(t)=e^{-\lambda t}$	Simple 'first order' model with a single coefficient, λ. This gives a risk of failure that is constant with time.
Weibull model[a]	$S(t)=e^{-(\lambda t)^{\alpha}}$	A more flexible model than the exponential model, with two coefficients, λ and α. For $\alpha>1$ the risk of failure increases with time; for $\alpha<1$ the risk of failure decreases with time.
Proportional hazard model[b]	$h_1(t)=\psi h_0(t)$	Models the relationship between two survival curves in terms of a single parameter ψ which quantifies the difference in risk of failure under two conditions. The model requires a control condition for comparison, so is most useful for studying the effect of changes in conditions, for example alternative packaging or higher storage temperature. The principal advantage of the model is that it permits comparison of failure risks when the form of the failure model is not known. The model is usually applied to data arising from constant monitoring of a complete population.

[a] For parametric models, the model form is given in terms of the survival function $S(t)$, which describes the probability of survival to time *t*.

[b] For the proportional hazard model, the model form is given in terms of hazard functions $h(t)$, which gives the probability of failure at time *t* given that an RM unit has survived to time *t*.

5.7 Commutability assessment

5.7.1 General considerations

5.7.1.1 The commutability of an RM relates to the ability of the RM, characterized by one measurement or testing 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 important for certain classes and uses of RMs (e.g. RMs for in vitro diagnostics).

NOTE 1 Demonstration of commutability is usually not required when the measurement 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 RM is obtained from sources and handled the same as the samples, for example, matrix RMs.

5.7.1.2 Where the RMP warrants that a RM is appropriate for a particular intended use which requires a commutable material, the RMP should undertake an assessment of commutability.

NOTE 1 ISO 15194, which is referenced by ISO 15195, requires evaluation of the commutability of certified RMs for in vitro diagnostic procedures.

NOTE 2 Further information is provided in Reference [37].

5.7.2 Commutability assessment for qualitative properties

5.7.2.1 Commutability is important for qualitative property RMs that are intended for use as positive or negative controls in a qualitative testing procedure. The general provisions of 5.1 apply to these materials.

5.7.2.2 Commutability assessment should include one or more of the following:

— consideration of all the physical, chemical and biological properties of the material that could reasonably affect commutability, including the particular characteristics certified and any processing or other treatment of the RM to improve stability, homogeneity or usage properties;

— review of published literature data on commutability of closely related materials;

— planned experiments to test commutability.

5.7.2.3 Experimental commutability assessment for qualitative property RMs is intended to provide experimental evidence that the material responds in the same way, under examination procedures that can be applied in conditions of intended use, as do typical test materials expected to share the same characteristics. An experimental commutability study should be undertaken:

— where consideration of the properties of the material indicates that absence of commutability could adversely affect use of the material;

— where required for compliance with applicable standards.

5.7.2.4 For characteristics for which there is only one possible examination procedure, it is sufficient to show that the RM responds as expected to that procedure. Where the RM is intended for use across multiple procedures, the response to commonly used procedures should be confirmed. Where different examination procedures operate on different principles, the RM's response to representative procedures operating on each different principle should be confirmed.

EXAMPLE Where a matrix RM is certified as containing a particular potentially allergenic food ingredient, examination procedures can operate on the principles of nucleic acid sequence detection by polymerase chain reaction methods, or protein detection by immunoassay or liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) methods. It is then important to confirm that the material responds correctly to representative nucleic acid and protein detection methods routinely applied for detecting the particular food ingredient.

5.7.3 Commutability statement

Where commutability information is required, the RMP should provide sufficient information for the end user to judge whether the material is appropriate for the specified use without further qualification or whether additional qualification by the end user is required before use. In particular, the certificate or associated documentation should make clear:

— whether commutability studies have been carried out;

— where a study has been carried out, the test methods for which the material has been shown to be commutable and any for which the material has been shown not to be commutable;

— any differences between the RM and routine test materials which are known to the RMP and which could reasonably reduce commutability for other test methods.

Annex A
(informative)

Guidance for DNA and protein reference materials

A.1 DNA reference materials

A.1.1 Value assignment of qualitative DNA reference materials

A.1.1.1 In principle, two kinds of values can be assigned to a DNA RM: one is the determination of the sequence of nucleobases – A (adenine), C (cytosine), G (guanine) and thymine (T) – in a relevant portion of the target DNA strand, the other the identity of the species from which the RM is produced.

A.1.1.2 DNA sequencing is a technique based on obtaining and evaluating quantitative measurement signals. These signals are evaluated and a conclusion about the nominal property in form of a DNA sequence is taken. Guidance about the evaluation of the quality of sequencing data can be found in ISO 20397-2.

A.1.1.3 The value assignment of DNA RMs certified for a nominal property relies on:

— controlled origin and purity of the raw materials used for processing;

— value assignment using DNA sequencing or other DNA-based techniques (e.g. polymerase chain reaction (PCR)).

A.1.2 Controlled origin and purity

A.1.2.1 The RMP is encouraged to ensure that the raw material used for RM processing has the assumed identity by acquisition of objective evidence, for example:

— genomic DNA extracted from a specific microorganism strain obtained from culture collections together with a certificate of analysis confirming the identity of the material;

— DNA fragments selected for cloning, including well controlled cloning conditions (e.g. single-colony cultivation by using high-copy vectors from the same incompatibility group);

— animal or plant tissue samples of known identity and origin (e.g. morphological identification by a taxonomist of the species, DNA sequence known and genetic identification performed).

A.1.2.2 The purity of the RM can play an important role and should be assessed. The impact of potential DNA contaminants for the intended use of the RM should be assessed, for example remaining traces of genomic DNA from host bacterial cells or traces of ribonucleic acid (RNA) molecules can induce a bias in the UV absorbance-based DNA quantification. Likewise, impurities can impact the outcome of measurements aiming at the verification of the identity.

A.1.3 Application to other nucleic acid material

A.1.3.1 The principles in A.1.1 and A.1.2 may also be applied to other nucleic acid RMs (e.g. RNA).

A.2 Protein reference materials

A.2.1 Value assignment of qualitative protein reference materials

Protein identification is the process by which a protein's primary amino acid sequence is determined or the distinct name (or accession number) of a particular protein of interest is determined based on its constituent amino-acid residues linked to primary DNA sequences of known gene holotypes (type specimen) of the organism. Identity is the nominal property of interest for protein identification. Protein analysis can be used for the identification, quantification, purity assessment, expression, post-translational modification determination, induction and turnover (degradation and resynthesis) of groups or individual proteins. Protein identification is an essential step in the field of proteomics; however, without identifying the proteins known to be critically involved in the system under investigation, it is not possible to investigate the biological explanation of the system of interest. The value assignment of protein identity relies on the determination of a proteins' primary sequence of amino acid residues.

A.2.2 Determination of amino acid sequence

A.2.2.1 The amino acid sequence in a single protein standard should be determined by demonstrably reliable techniques such as high-resolution mass spectrometry (MS) combined with Edman degradation. Where appropriate, determination of the primary amino acid sequence should be compared with highly annotated protein databases acquired from primary DNA sequences.

NOTE 1 Examples of highly annotated protein databases include Entrez Protein database[38], Reference Sequence (RefSeq) database[39] and UniProt[40], consisting of Swiss-Prot[41] and its supplement, TrEMBL[42].

NOTE 2 Highly annotated typically refers to the database having a description of the protein function, domain structure, post-translational modifications, variants and isoforms with a minimal level of redundancy. There are several examples of these databases, including Bos taurus (bovine), Mus musculus (mouse), Saccharomyces cerevisiae (baker's yeast) and Escherichia coli, with the best example being Homo sapiens (human).

A.2.2.2 Additional verification of a protein primary amino acid sequence can be attained by peptide mass fingerprinting. Regardless of the technique or techniques used for amino acid analysis to assign the primary protein sequence of the RM, a coverage factor to the known sequence should be demonstrated and a reference to the database and version of the known sequence should be included in all RM documentation. The coverage factor should be comparable to that of a suitable reference protein that is co-analysed with the sample.

A.2.3 Identification of a single protein or proteins in complex materials

A.2.3.1 The protein identification in complex materials should be carried out by demonstrably reliable techniques such as LC-MS/MS, followed by spectral identification by sequence database searching. Documentation of all necessary sample preparation steps used in the protein identification (e.g. solubilization media, S-S bond reduction, free S residue treatment, proteolytic enzyme) should be included. Additionally, the search algorithm and software version number and/or release date used for identification (e.g. Byonic, Comet, Mascot, MS Amanda, MS-GF+, Sequest), the source and version of protein database used, the database search parameter criteria for identifications (e.g. mass range, fragmentation mode, mass analyser, resolution), the number of missed and non-specific cleavages permitted, a list of all modifications (static and variable) considered and any additional parameters used in the identification process for RM assignment should be documented. The criteria (e.g. threshold score or factor criteria used for accepting spectra as sufficient basis for protein identification) should be documented. The false discovery rate (FDR) [43-45] at the peptide and protein levels, including any additional statistical analyses that estimate a measure of identification certainty or allow a determination of the FDR (e.g. the results of decoy searches or other statistical and computational approaches), should also be documented.

A.2.3.2 Identification based on more than one search algorithm impacts the reliability of the identification assignment and increasing number of algorithms can increase the confidence of the value assigned identity of a protein. Additional verification of a protein identity in complex materials can be accomplished by LC-MS/MS with spectral identification by de novo sequencing and/or spectral identification by hybrid methods.

A.2.3.3 Due to the updating of both protein databases and software algorithms used for protein identification, it is recommended that data used for the value assignment of protein identification in RMs, including raw data and processed data as well as both experimental and technical metadata, should be archived in an appropriate publicly available data repository (e.g. PRIDE, MassIVE, Peptide Atlas, Panorama).

Annex B
(informative)

Expressing confidence in qualitative values

B.1 General considerations for expressing confidence in qualitative values

B.1.1 The assignment of value to a qualitative property of a RM generally is surrounded by uncertainty, similar to the assignment of value to quantitative properties. For qualitative properties, however, the prevailing practice is to express confidence in such assignment, rather than expressing uncertainty. Such expression of confidence can be quantitative or qualitative. When it is qualitative, it typically uses an ordinal scale of levels of confidence.

B.1.2 The qualitative evaluation should be based on explicit criteria that are specifically appropriate and unambiguous for application in the relevant context and which should be stated in the corresponding certificate.

EXAMPLE A CRM comprises leaves of *Ginkgo biloba*. The uncertainty associated with the identification of the biological species was expressed relative to the following ordinal scale[46]:

— MOST CONFIDENT (Have very well-supported and well-resolved phylogeny and/or multiple diagnostic nucleotides differentiating species from closest relatives; have data from multiple samples of both an inclusivity and exclusivity panel; data from multiple independent gene regions agree).

— VERY CONFIDENT (Have reasonably well-supported and well-resolved phylogeny and/or a few diagnostic nucleotides differentiating species from close relatives; have data from multiple samples of both an inclusivity and exclusivity panel; data from one gene or data from multiple independent gene regions agree).

— CONFIDENT (Have reasonably well-supported and well-resolved phylogeny and/or one or a few diagnostic nucleotides differentiating species from close relatives; have data from a few samples of both an inclusivity and exclusivity panel; data from one gene, or data from multiple independent gene regions generally agree).

— AMBIGUOUS (Have a poorly supported and poorly resolved phylogeny and/or no diagnostic nucleotides differentiating species from close relatives; have data from a few or multiple samples of both an inclusivity and exclusivity panel; data from one gene or data from multiple independent gene regions generally disagree).

B.2 Performance measures for qualitative property tests

B.2.1 A range of performance measures are in use for describing the performance of test methods (including expert assessments) which return only a nominal result[47]. These can be described by reference to an experimental assessment in which a number of items with known nominal values are tested and a result recorded as one of two outcomes (here denoted positive and negative). Table B.1 summarises the possible outcomes and counts in such an assessment, assuming a test with only two possible outcomes. Table B.2 describes some common performance indicators that can be derived from the counts in Table B.1.

NOTE 1 The performance measures in Table B.2 are most commonly applied to binary test methods, for which "positive" and "negative" represent, for example, presence and absence of a single feature. They can, however, be used to indicate the presence of a particular value among a larger set of mutually exclusive possibilities (e.g. a particular biological or chemical species), so that a correct classification is regarded as a true response and misclassification is counted as a false response. This permits application to cases where more than two values are possible.

NOTE 2 Subclause B.3 gives information on measures of agreement for nominal properties for which a graduated degree of agreement can be established.

B.2.2 False response counts (n_{FP} , n_{FN}) for effective test methods are usually low and can, by chance, be zero even when the underlying rate is appreciable. To avoid severe underestimation of false response rates, it can be useful to substitute an upper confidence limit for an observed count or to apply other methods of inference. For example, one simple procedure, sometimes called "pseudocounting", is to add one to each observed count in a study before calculating proportions.

NOTE The addition of one to each observed count can be justified on Bayesian grounds. Given a uniform prior for the probability p_F of a particular false response, to be estimated from an observed counts n_F of false responses and a total of $N = n_F + n_2 + \ldots n_n$ responses for the population of interest, it can be shown that the mean of the posterior distribution for p_F is $(n_F + 1)/(N + n)$; that is, the proportion calculated after addition of one to each count. Although positively biased for low n_F, this is often a more useful estimate of false response rate when zero counts are common.

B.2.3 None of the performance measures in Table B.1 can be directly interpreted as probabilities that a particular nominal value is correct (or incorrect). Several can, however, be interpreted as (estimates of) conditional probabilities which can inform a quantitative (probabilistic) indication of confidence in a nominal value. Two such indications are described in C.2.

B.2.4 In addition to simple counts of binary outcomes, the reliability of procedures for value assignment for nominal properties with multiple possible outcomes, including ordered scales, can be characterised using measures of agreement. Some examples of such measures are listed in Table B.3. Formulae are given in the associated bibliographic references.

NOTE See Reference [18] for guidance.

Table B.1 — Experimental counts for assessment of nominal properties

Test result	Known or confirmed test item condition		Test result totals
	Positive	Negative	
Positive (P)	n_{TP}	n_{FP}	$n_{P,obs}$
Negative (N)	n_{FN}	n_{TN}	$n_{N,obs}$
Item totals	n_P	n_N	N
NOTE The test method is assumed to provide one of two possible results, "positive" or "negative", which is either correct ("true") or incorrect ("false")			

Table B.2 — Performance characteristics for binary test methods

Performance characteristic	Formula[a]	Description
False positive rate	$\frac{n_{FP}}{n_{TN} + n_{FP}} = \frac{n_{FP}}{n_N}$	Proportion of negative test items reported as positive. Lower is better.
False negative rate	$\frac{n_{FN}}{n_{TP} + n_{FN}} = \frac{n_{FN}}{n_P}$	Proportion of positive test items reported as negative. Lower is better.
Sensitivity	$\frac{n_{TP}}{n_{TP} + n_{FN}} = \frac{n_{TP}}{n_P}$	Proportion of positive test items correctly reported as positive. Higher is better.
Specificity	$\frac{n_{TN}}{n_{TN} + n_{FP}} = \frac{n_{TN}}{n_N}$	Proportion of negative test items correctly reported as negative. Higher is better.
Positive predictive value	$\frac{n_{TP}}{n_{TP} + n_{FP}} = \frac{n_{TP}}{n_{P,obs}}$	Proportion of positive results that are correct. Higher is better.
[a] The formulae provide estimates of underlying proportions or indices based on the experimental counts in Table B.1.		

Table B.2 *(continued)*

Performance characteristic	Formula[a]	Description
Negative predictive value	$\frac{n_{TN}}{n_{TN}+n_{FN}} = \frac{n_{TN}}{n_{N,obs}}$	Proportion of negative results that are correct. Higher is better.
Efficiency	$\frac{n_{TP}+n_{TN}}{n_{TP}+n_{TN}+n_{FP}+n_{FN}}$	Proportion of all results that are correct. Higher is better.
Youden Index	Sensitivity (%) + Specificity (%) − 100	Summary performance figure for convenience. Higher is better.

[a] The formulae provide estimates of underlying proportions or indices based on the experimental counts in Table B.1.

Table B.3 — Examples of performance characteristics for qualitative test methods

Performance characteristic or procedure	Applicability			Range of values their interpretation	References
	No. of value assignment procedures[a]	No. of test items[b]	No. of categories		
Cohen's kappa	2	≥ 2	≥ 2[c]	−1 to +1 (higher is better)	[48],[49]
Fleiss' kappa	≥ 2	≥ 2	≥ 2[c]	−1 to +1 (higher is better)	[48],[50],[51]
Attribute agreement analysis[c]	≥ 2	≥ 2	≥ 2[c]	0 % to 100 % (higher is better)	ISO/TR 14468
Kendall's coefficient of concordance	≥ 3	≥ 2	≥ 2 (ordinal scale)	0 to 1 (higher is better)	[49],[52]
CATANOVA	≥ 2	≥ 2	≥ 2	0 to 1 (lower is better)[d]	[28]
ORDANOVA (for ordinal ANOVA)	≥ 2	≥ 2	≥ 2 (ordinal scale)	0 to 1 (lower is better)[d]	[29],[53]

[a] The number of procedures can include the set of reference values, if known. That is, a study in which a single nominal value assignment procedure is checked against a set of reference values is considered to be a system of two value assignment procedures.

[b] Although most procedures can be applied to small numbers of test items, the results become more informative as the number of items increases.

[c] Although applicable to multiple possible responses, the procedure records results as either in agreement or not; that is, scoring is binary.

[d] Kendall's coefficient of concordance uses the rank order difference between different values on an ordinal scale as a measure of disagreement.

NOTE Like analysis of variance, ORDANOVA provides quantitative measures of variation (disagreement); the recommended measures are normalised to the range 0 to 1. Higher values indicate less agreement. Interpretation uses tests for significant difference based on simulation.

B.3 Quantitative indicators of agreement or similarity for nominal values

B.3.1 Introduction to indicators of agreement

Some nominal values are assigned by comparing a large number of features with a reference. For example, a biological species could be identified from a DNA sequence in which many base pairs in an experimentally determined DNA sequence match a reference sequence for the designated species. In other cases, a material could be identified from a combination of qualitative and quantitative observations. Where this is the case, measures of agreement between the set of expected values and the observed values can provide an indication of confidence in the assigned value. Two measures are described briefly in B.3.2.

B.3.2 The Gower Index

B.3.2.1 The Gower Index of Similarity[54], often abbreviated to "Gower Index" or "Gower Similarity", has been applied to the characterisation of RMs with qualitative properties as described in ISO/TR79. It has the advantage of accommodating binary, ordinal scale and quantitative information. It is calculated as in Formulae (B.1) and (B.2):

$$g = \frac{\sum_{i=1}^{n} w_i \, g_i}{\sum_{i=1}^{n} w_i} \tag{B.1}$$

$$g_i = 1 - \frac{|x_i - x_{0,i}|}{r_i} \tag{B.2}$$

where

g is the Gower similarity coefficient defined by Formula (B.1);

g_i is a similarity measure for the i^{th} characteristic;

w_i is the weight allocated to the i^{th} characteristic;

x_i is the observed value of the i^{th} characteristic;

$x_{0,i}$ is a chosen reference value for the i^{th} characteristic;

r_i is the range of values for the i^{th} characteristic.

B.3.2.2 The range r_i is chosen so that g_i is between 0 and 1. Weights w_i are assigned according to the problem. Values of g near 1 express close agreement with expected values; values of g near zero show poor agreement. The index can be used as a criterion for nominal value assignment or for conveying a closeness of match in a report or an RM document.

NOTE ISO/TR 79 gives further details of calculation and use of the Gower coefficient in a RM context.

B.3.3 Nucleic acid or protein sequence match scores

Biological species are often identified or confirmed on the basis of a DNA or protein sequence match. For nucleic acid or protein sequences, measures of sequence similarity are used to indicate the degree of similarity between two sequences. Usually, the degree of similarity is quantified by a score. A variety of different scoring systems have been used[55]. Interpretation often uses an estimated probability of a chance match with the same or better score. This can be used, with care (see 5.3.3.3 and 5.3.3.4), to give a quantitative indication of confidence in the assigned species.

NOTE Sequence similarity can be reported as "percent identity", the percentage of characters that match between two sequences.

B.4 Selected expressions of confidence in nominal values

B.4.1 Likelihood ratio

B.4.1.1 A likelihood ratio can be used to convey the "weight of evidence" that is provided by a particular test result[56,57]. The likelihood ratio gives the increase in odds in favour of one hypothesis over another that results from new information, such as a positive test result. In the case of a nominal value, the hypotheses of interest are H1: the assigned value is correct and H2: the assigned value is not correct. Assuming that a test

has provided a positive result on a candidate nominal property RM and that the assigned value is accordingly positive, the likelihood ratio $L_{H1:H2}$ is given by Formula (B.3).

$$L_{H1:H2} = \frac{1-p(N|P)}{p(P|N)} = \frac{p(P|P)}{p(P|N)} \quad \text{(B.3)}$$

where

$p(N|P)$ is the estimate of false negative rate in Table B.1 (here written as the conditional probability of a negative response given a positive test item);

$p(P|N)$ is the relevant false positive rate, again written as a conditional probability;

$p(P|P)=1-p(N|P)$ is the probability of a positive result given a positive test item, that is, a true positive rate. For this simple case, the likelihood ratio is the true positive rate divided by the false positive rate.

B.4.1.2 A likelihood ratio of the form shown in Formula (B.3) can take any positive value. Values above 1 indicate support for hypothesis H1; values below 1 indicate support for hypothesis H2. For tests used in RM value assignment, false response rates are normally low for both false negatives and false positives. Likelihood ratios are accordingly expected to be in excess of 1 000. For some DNA identification applications, which involve a combination of multiple independent tests, likelihood ratios can exceed 10^9.

NOTE A high likelihood ratio does not guarantee a correct result or describe the probability that the test result is correct. Rather, it shows how well the test result supports the assigned value considered against a particular alternative. A high likelihood ratio provides strong evidence in favour of the reported value. If information is available which shows that the likelihood ratio is not applicable, then the likelihood ratio cannot be used.

B.4.2 Posterior probability

B.4.2.1 A posterior probability is a concept arising in Bayesian statistics. It can be used to give a direct indication of the probability that an assigned nominal value is correct[57].

B.4.2.2 A posterior probability for a nominal value is formed from a combination of prior information on the incidence or likelihood of a particular nominal value occurring in a candidate material, combined with additional information which will (in RM value assignment) generally act to increase the calculated probability. The framework for this process is Bayes' Theorem.

For the purpose of nominal value assignment, Bayes theorem can be expressed according to Formula (B.4).

$$p(A|E) = P(A)\frac{P(E|A)}{P(E)} \quad \text{(B.4)}$$

where

$p(A|E)$ is the probability associated with the assigned value following evidence (usually a test result or results) E; $p(A|E)$ is known as the posterior probability;

$p(A)$ is the probability associated with the assigned value prior to the value assignment exercise, that is, the prior probability of A;

$p(E|A)$ is the probability of obtaining the evidence (or test result) E from a material for which the nominal property of interest has the value A;

$p(E)$ is the probability of obtaining the test result E irrespective of the value of the nominal property of interest.

NOTE 1 A conditional probability of the form $p(x|y)$ is usually read as "the probability of event x given event y" or simply "the probability of x given y".

NOTE 2 $p(E)$ is evaluated for a population of interest, which may have high or low proportions of materials for which the property of interest has value *A*. Sometimes it is useful to write $p(E)=p(E|A)+p(E|\bar{A})$, where $\bar{A}$ denotes "Not *A*". This gives a clear indication of the importance of instances in which a test result arises in the absence of material with the value *A*, i.e. false positives.

NOTE 3 Formula (B.4) is sometimes referred to as "Bayes' Rule".

NOTE 4 The fraction $p(E|A)/p(E)$ in Formula (B.4) is sometimes referred to as the likelihood, so that "the posterior probability is the prior probability multiplied by the likelihood".

Formula (B.4) may in principle be applied recursively to a series of independent test results, substituting the posterior from one calculation into the prior for the next to obtain a posterior probability after an accumulation of multiple independent tests.

There are many potential pitfalls in application of Formula (B.4). These include, but are not limited to the following:

— Difficulty in assigning a reliable prior probability $p(A)$.

— Difficulty in obtaining a reliable probability $p(E)$ for a given population.

— Difficulty in defining the population of interest and the proportion of that population which can, correctly or incorrectly, generate the test result *E*.

— Strong sensitivity to lack of independence among test results or observations. For example, in spectroscopy, separate signals are rarely completely independent of one another and may co-occur for related compounds.

B.4.2.4 The potential pitfalls listed in B.4.2.3 can lead to errors of several orders of magnitude in estimating posterior probabilities. Successful application of Formula (B.4) therefore requires:

— sound information on the population of materials of interest;

— accurate information on false positive and false negative rates for different parts of the population;

— well-established false response rates for all relevant tests;

— full information on any association (lack of independence) among results for different tests;

— staff with suitable qualifications and experience in the application of Formula (B.4).

B.4.3 Entropy

B.4.3.1 The uncertainty associated with an assignment of value to a nominal property is most thoroughly represented by a probability distribution on the set of possible values that the property can take.

EXAMPLE When determining the identity of a nucleobase at a particular locus in a DNA strand, the sequencer could yield the following probabilities: Distribution (a) 0,02 for A, 0,00 for C, 0,01 for G and 0,97 for T conveys greater uncertainty than distribution (b) 0,04 for A, 0,09 for C, 0,06 for G and 0,81 for T..

B.4.3.2 Entropy quantifies the extent to which the unit of probability is spread out over the set of values that the property can take. If this set is {a, b, c, ...} and the corresponding probabilities are $\{p_a, p_b, p_c, \ldots\}$ (non-negative numbers adding to 1), then the entropy of this probability distribution is $H=-(p_a\times\ln(p_a)+p_b\times\ln(p_b)+p_c\times\ln(p_c)+\ldots)$, where "ln" denotes the natural logarithm (base e = 2,718 282...). Since the larger the H the greater the uncertainty, the entropy can be a useful expression of the uncertainty associated with an assignment of value to a nominal property (see Example E6 in Reference [58]).

EXAMPLE For the first case in the example in B.4.3.1, $H=-(0{,}04\times ln(0{,}04)+0{,}09\times ln(0{,}09)+0{,}06\times+0{,}81\times ln(0{,}81))=0{,}685$; for the second case, $H=-(0{,}02\times ln(0{,}02)+0{,}00\times ln(0{,}00)+0{,}01\times ln(0{,}01)+0{,}97\times ln(0{,}97))=0{,}154$, where the fact that $0\times ln(0)=0$ has been used.

B.4.3.3 The entropy of a continuous probability distribution with probability density function p is the integral of $-p\times\ln(p)$ over the set of real numbers where $p>0$. For the Gaussian (or normal) distribution with mean μ and standard deviation σ the entropy is $H=\ln(\sigma)+0{,}5\times\ln(2\pi e)$. Therefore, σ is proportional to $\exp(H)$, which shows, albeit in a particular case, that the larger the H the larger the σ and also suggests that, in general, $\exp(H)$ may be regarded as an analogue of the standard uncertainty, up to a multiplicative factor.

EXAMPLE Given the refractive index and the mass fractions of the major oxides of a glass fragment recovered in a forensic investigation, a classifier yielded the following probabilities for its provenance: 0,36 (contemporary building window), 0,56 (old building window), 0,08 (vehicle window), 0,00 (bottle), 0,00 (tableware) and 0,00 (automobile headlamp). The entropy of this distribution is $-(0{,}36\times\ln(0{,}36)+0{,}56\times\ln(0{,}56)+0{,}08\times)=0{,}895$. The "equivalent" standard uncertainty would be $exp(0{,}895)/exp(0{,}5\times\ln(2\pi e))=0{,}592$ (see Example 29 in Reference [58]).

Annex C
(informative)

Statistical procedures

C.1 Chi-squared test for a two-way contingency table

C.1.1 Purpose

The following procedure provides a test for significant differences in the proportion of non-compliant results among different groups of results. It can be applied for:

— testing whether RM units differ significantly, following a homogeneity study in which replicated tests are performed on each RM unit (see 5.5);

— testing for evidence of an effect of time or some other condition (such as temperature), following a stability study in which multiple RM units are exposed to different conditions and subsequently tested (see 5.6).

NOTE The procedure described here is appropriate for a single grouping factor, which can be either RM unit (in the case of a homogeneity study) or a single treatment condition, such as time or temperature (in the case of a stability study).

C.1.2 Data requirement

The procedure requires the following:

a) For a homogeneity study: a table of counts of compliant and noncompliant (or positive and negative) results, by RM unit.

b) For a stability study: a table of counts of compliant and noncompliant (or positive and negative) results, by exposure condition.

Example layouts are shown in Table C.1 and Table C.2.

NOTE The table layouts are shown as horizontal for convenience of layout in this document. For a chi-squared test, the orientation of the table is not important; rows and columns can be interchanged as convenient.

C.1.3 Procedure

Denote the number of table columns as n_C and the number of rows as n_R. Denote the count in the table cell at row i and column j of the table as n_{ij}. Then do the following:

a) Calculate the total r_i for each row according to Formula (C.1).

$$r_i = \sum_{j=1}^{n_C} n_{ij} \tag{C.1}$$

b) Calculate the total c_j for each column according to Formula (C.2).

$$c_j = \sum_{i=1}^{n_R} n_{ij} \tag{C.2}$$

c) Calculate the total number of observations N according to Formula (C.3).

$$N = \sum_{i=1}^{n_R} \sum_{j=1}^{n_C} n_{ij} \tag{C.3}$$

d) Calculate the expected count E_{ij} for each table cell according to Formula (C.4).

$$E_{ij} = \frac{r_i \cdot c_j}{N} \tag{C.4}$$

e) Calculate the chi-squared test statistic χ^2 according to Formula (C.5).

$$\chi^2 = \sum_{i=1}^{n_R} \sum_{j=1}^{n_C} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{C.5}$$

f) Calculate the degrees of freedom ν according to Formula (C.6).

$$\nu = (n_R - 1)(n_C - 1) \tag{C.6}$$

g) Obtain the critical value $\chi^2_{0,95,\nu}$, which is the upper 0,95 quantile of the chi-squared distribution with ν degrees of freedom, from a table of quantiles of the chi-squared distribution or from software.

h) Compare the test statistic χ^2 with the critical value $\chi^2_{0,95,\nu}$. If the test statistic χ^2 is greater than the critical value $\chi^2_{0,95,\nu}$, declare the test result as significant at the 95 % level of confidence.

For a homogeneity test, a significant test result indicates that the material is not homogeneous. For a stability test, a significant test result indicates that there are differences in failure proportion at different times.

NOTE The reference probability distribution for the test criterion defined in C.1.3(e) is approximately chi-squared when the counts in all the table's cells are 5 or larger. Where many counts are below 5, Fisher's exact test can be applied[27].

NOTE 5.6.2.3 gives additional guidance on the interpretation of stability tests using a chi-squared test.

Table C.1 — Example layout for basic homogeneity study

Test result	RM unit identifier						
	#1	#2	#3	…	#(j)	…	#(n_p)
Compliant	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	…	$n_{1,j}$	…	n_{1,n_C}
Non-compliant	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	…	$n_{2,j}$	…	$n_{2,NC}$
NOTE $n_{i,j}$ denotes the number of observations (compliant or non-compliant, respectively) in each cell of the table.							

Table C.2 — Example layout for basic stability study

Test result	Time point					
	#1	#2	…	#(j)	…	#(n_p)
Compliant	$n_{1,1}$	$n_{1,2}$	…	$n_{1,j}$	…	n_{1,n_C}
Non-compliant	$n_{2,1}$	$n_{2,2}$	…	$n_{2,j}$	…	$n_{2,NC}$
NOTE $n_{i,j}$ denotes the number of observations (compliant or non-compliant, respectively) in each cell of the table.						

C.2 Confidence limit for proportion of non-compliant RM units

C.2.1 Purpose

The following procedure provides an upper 95 % confidence limit for the proportion π_{NC} of non-compliant RM units.

C.2.2 Data requirement

The procedure requires the following information:

— the number n_h of RM units tested in a homogeneity study;

— the number n_{NC} of RM units found to be non-compliant in the study;

— the required level of confidence $1-\alpha$; for a 95 % upper limit, $\alpha = 0,05$;

— the $(1-\alpha)$ quantile, $z_{1-\alpha}$, for the normal distribution; for a 95 % upper limit, $z_{1-\alpha} = 1,645$.

C.2.3 Procedure

The procedure is based on the Wilson scoring interval[25]:

a) Calculate the estimated proportion defective, $\hat{p}_{NC}$, according to Formula (C.7).

$$\hat{p}_{NC} = \frac{n_{NC}}{n_h} \tag{C.7}$$

b) Calculate the upper 95 % confidence limit for proportion non-compliant, $L_{p_{NC,u}}$, according to Formula (C.8).

$$L_{p_{NC,u}} = \frac{\hat{p}_{NC} + \frac{z_{1-\alpha}^2}{2n_h} + z_{1-\alpha}\sqrt{\frac{\hat{p}_{NC}(1-\hat{p}_{NC})}{n_h} + \frac{z_{1-\alpha}^2}{4n_h^2}}}{1 + \frac{z_{1-\alpha}^2}{n_h}} \tag{C.8}$$

NOTE Confidence limits for other levels of confidence can be calculated by replacing $z_{1-\alpha}$ with the corresponding quantile of the normal distribution.

EXAMPLE A homogeneity test checks 120 RM units, of which two are found to not respond to the test as expected and are marked as noncompliant. Then:

$$\hat{p}_{NC} = \frac{2}{120} = 0,016\,7$$

$$L_{p_{NC,u}} = \frac{0,016\,7 + \frac{1,645^2}{2\times 120} + 1,645\sqrt{\frac{0,016\,7(1-0,016\,7)}{120} + \frac{1,645^2}{4\times 120^2}}}{1 + \frac{1,645^2}{120}} = 0,049$$