Anal Bioanal Chem (2011) 400:1729–1741
DOI 10.1007/s00216-011-4776-y

REVIEW

Uncertainty of measurement and conformity assessment: a review

Elio Desimoni · Barbara Brunetti

Received: 27 October 2010 / Revised: 1 February 2011 / Accepted: 4 February 2011 / Published online: 25 February 2011

Abstract The uncertainty of measurement is the key indicator of the quality of any experimental result. Proper consideration of this uncertainty is imperative when testing a sample against legal/compositional limits. This task can be quite challenging when the entity measured in the investigated sample is so close to the limit that its uncertainty, however estimated, critically affects decision-making. This explains the many literature contributions discussing the problem. Even though some of the most authoritative organisations have issued specific guidelines aimed at assisting the staff involved in such measurements, several aspects of conformity testing are still debated in the literature. In this review, after a short outline of existing information, somewhat more detailed insight is given into the guidelines of ASME, ISO, and Eurachem/CITAC, because they are the most useful tools for operators of testing and calibration laboratories. Some aspects of Council Directive 96/23/EC are also discussed. Insight into the contents of the mentioned documents enables emphasis of analogies and discrepancies.

Keywords Conformity testing · Decision rules · Limiting values · Specification limits · Guard band · Uncertainty of measurement

Introduction

It is well known that, when reporting the result of a measurement of a physical quantity, it is mandatory to give a quantitative indication of its quality, so that the user of the result can assess its reliability [1]. Such an indication is represented at best by the measurement uncertainty (MU), the value associated with the result of a measurement that characterises the dispersion of the values that could reasonably (e.g. with a given probability/confidence level) be attributed to the measurand [1]. As emphasized by the ISO Guide to the Expression of Uncertainty in Measurement (GUM), without a clear indication of their uncertainty, measurement results cannot be compared either among themselves or with reference values given in a specification or standard [1].

Unfortunately, when dealing with measurements aimed at evaluating conformity with some specification, the matter becomes quite complex when the measured entity in the sample under investigation (e.g. the concentration in chemical analyses) is so close to the specification that the MU, anyhow estimated, critically affects decision-making. This explains the uninterrupted appearance of contributions devoted to discussion of the multi-faceted aspects of considering the MU when assessing conformity to legal or compositional limits [2–32]. As can be seen, these papers, listed in chronological order, span the last fifteen years. Noticeably, a few of them appeared even after some of the most authoritative organisations issued specific guidelines aimed at assisting the staff involved in such measurements [33–36]. This is probably indicative of a still ongoing debate.

In this review, after a short outline of existing literature information a somewhat more detailed insight is given into the guidelines of ASME [33], ISO [34], and Eurachem/CITAC [35], because they are the most useful tools for operators of testing and calibration laboratories. Of course, this paper is not aimed exhaustively at presenting the three standards, a task that would obviously require much more

E. Desimoni (✉) · B. Brunetti
DiSTAM, Università degli Studi di Milano,
Via Celoria 2,
20133 Milano, Italy
e-mail: elio.desimoni@unimi.it

extensive types of presentation. Its only objective is to assist a reader in approaching those guidelines and comparing the basic concepts presented therein. Council Directive 96/23/EC [36] is also mentioned and partly discussed. Insight into some of the basic contents of these four last documents [33–36] enables emphasis of a few analogies and discrepancies.

Interpreting analytical results affected by measurement uncertainty against limiting values

Conformity[1] testing is the systematic examination of the extent to which an entity conforms to a specified criterion [34]. A specification for a measurable characteristic (for example, the concentration in chemical analysis) is usually formulated as a single limiting value, e.g. an upper *or* a lower limiting value, LV_U or LV_L, respectively, or as a set of limiting values, e.g. both an upper *and* a lower limiting value. The term specification limit, SL, is also used in place of limiting value.

Most frequently, when dealing with a set of limiting values, permitted values of the characteristic are those falling within the LV_L–LV_U interval. But, in some cases, permitted values are those falling outside that interval. An example of this last situation is that relevant to some inflammable compounds [13]: if their concentration in air is below a given LV_L, the gaseous mixture cannot burn or explode whereas if it is above the LV_U, the mixture can burn but it cannot explode. Within the two limits the mixture explodes.

Several papers have presented the very basic aspects of interpreting how experimental results, being affected by MU, should be interpreted against some specification limits. The problem is schematised by more or less detailed figures in which different measurement results, with their MU interval, are compared with one LV or with a set of LVs [5, 7, 10, 16, 20, 26, 27, 31, 37–39]. The uncertainty interval is estimated according to a given confidence level, usually 95% (see the next section). Most frequently, the problem is presented as in Fig. 1, or as in its top half. Four possible experimental situations are recognisable at each LV. Occasionally, an additional situation is added in which the measurement result coincides with a limit [7, 10, 37]. In one case, eight different situations are considered [16]. But the four situations A–D of Fig. 1 allow any possible reasoning. By limiting the attention at the upper limiting value only, one can easily argue that in case A the product

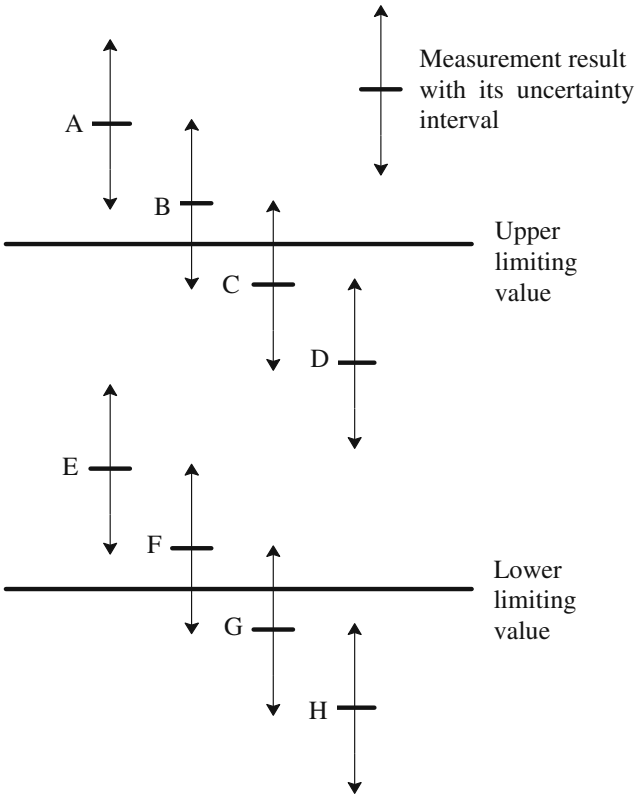

Fig. 1 Assessing conformity when the measurement result is more or less close to a higher or lower limit. In these figures, as usual, the permitted values of the characteristic are those falling within the LV_L–LV_U interval

does not comply with the specification, because the whole uncertainty interval is above the limit, whereas in case D the product complies with the specification, because the whole uncertainty interval is below the limit. Of course, these two cases do not pose any problem of decision making at the selected confidence level.

In the two remaining cases, B and C, the uncertainty interval encompasses the LV_U, so knowledge of the measurement result does not enable any decision making: the result lies in the so-called uncertainty range. Case B does not allow statement of conformity at the chosen level of confidence (for example, 95%) even if non-conformity is more probable than conformity. The opposite applies in case C, in which conformity is more probable that non-conformity. Then, cases B and C are those requiring further investigation. The first possibility is that allowed by using a measurement method precise enough to reduce the MU interval at the level necessary to move from case B to case A or from case C to case D. This solution is not always possible, and usually implies a substantial rise of analysis cost and time. Alternatively, one can apply the two-stage procedure suggested by ISO 10576–1 [34] (see the section dealing with that standard). Again, additional measurements are necessary so that the cost and time of the analysis are

[1] In many of the references cited in this paper, the word "compliance" is used as a synonym of conformity. Strictly speaking, compliance indicates the action of making something conform or fulfilling a regulatory requirement.

accordingly increased. Some authors have suggested that, in cases such as B and C, stating conformity or non-conformity with a level of confidence lower than 95% is better than nothing [7, 10, 37]. However, such a possibility does not always appear realistic, as in the case of court cases in which the conformity or non-conformity statement must be "beyond reasonable doubt" [3].

Of course, the discussion about a lower limiting value (half bottom of Fig. 1) mirrors that detailed above.

It has also been reported that, even at present, it is possible that some specification makes no reference to properly considering the effects of MU on the assessment of conformity. In these cases "… it may be appropriate for the user to make a judgement of conformity, based on whether the test result is within the specified limits with no account taken of the uncertainty. This is often referred to as a *shared risk*, since the end-user takes some of the risk that the product may not meet the specification after being tested with an agreed measurement method" [37].

Finally, recent papers, when examining uncertain cases such as B and C in Fig. 1, also had the objective of evaluating the effect of MU on producer's and user's risk (usually associated to type I and, type II errors, respectively) in classification and conformity assessments [32].

Which uncertainty?

Nowadays, the term "uncertainty of measurement" is definitely used to mean the expanded uncertainty, U, obtained by multiplying the combined standard uncertainty, u_c (sometimes reported as $u_c(y)$, where y is the estimate of the measurand Y), by the coverage factor, k [1, 2, 7, 16, 22, 33–35, 37–39]. The intended purpose of U is to provide an interval around the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand [1].

The combined standard uncertainty is the total uncertainty of a measurement result estimated by properly combining all the uncertainty components affecting the whole experimental procedure. Whenever the procedure includes the sampling of the material under examination, it is mandatory considering the uncertainty of sampling among all the other uncertainty components (see for example Refs. [13–15, 34, 35, 39–42]). However, it should be also noted that the combined standard uncertainty is an estimated standard deviation relying on the assumption that no source of uncertainty has been neglected or overlooked and that, consequently, is itself affected by a more or less significant uncertainty. Also the GUM (section G 1.2) emphasizes that the value of the expanded uncertainty is at best only approximate [1].

The coverage factor is a multiplier chosen on the basis of the desired level of confidence to be associated with the interval defined by $U=k\cdot u_c$. Most frequently, k is in the range 2 to 3 [1]. When the normal distribution applies and u is a reliable estimate of the standard deviation of the measurand, $U=2\cdot u_c$ defines an interval having a level of confidence of approximately 95% (more exactly, a level of confidence of 95.45%), and $U=3\cdot u_c$ defines an interval having a level of confidence of approximately 99% (more exactly, a level of confidence of 99.73%). However, some aspects relevant to the concept of uncertainty still deserve specific comments.

Notwithstanding the detailed and authoritative documents intended to explain the meaning of uncertainty, it was noticed that surprisingly invalid MUs were sometimes provided, often labelled as "standard deviation" [18]. Examples of such erroneous estimates are an uncertainty resulting from calibration only, a repeatability standard deviation and a linearity of some calibration curve.

Moreover, using $k=2$ or 3 can no longer be accepted if the combined uncertainty has too few degrees of freedom [1, 19, 26]. If the effective number of degrees of freedom, ν_{eff}, is too low (for example, fewer then six according to Ref. [43], fewer than 30 according to Ref. [19]) the Student-t distribution is the most appropriate (approximate) choice for associating a level of confidence with U. It is known that ν_{eff} can be estimated by use of the Welch–Satterthwaite formula [1]. As underlined by the GUM, the experimental standard deviation of the mean of as many as thirty repeated observations of a measurand described by a normal distribution has itself an uncertainty of approximately 13% [1].

It is also possible that the reported uncertainty data look questionable. In these situations, the Horwitz equation [44–46] can sometimes provide a more realistic view [18, 19]. The equation is usually presented as:

$$RSD\% = 2^{(1-0.5\cdot\log(mf))} \quad (1)$$

Where $RSD\%$ is the among-laboratory relative standard deviation and mf is the analyte mass fraction (e.g. $mf=10^{-6}$ means mg kg^{-1}). The equation describes the precision performances of a measurement method without regard to the nature of the analyte, the type of test material, the applied analytical technique, and the complexity of the procedure. Equation 1 is presented in Fig. 2. Acceptable performances usually provide variability values within one-half to twice the predicted $RSD\%$ [45]. Within-laboratory values are expected to be one-half to two-thirds of the among-laboratory values [46]. Even if significant deviations from the estimates obtained by the Horwitz equation are possible (Ref. [47] and references cited therein) nevertheless, the equation is still an acceptable basis for

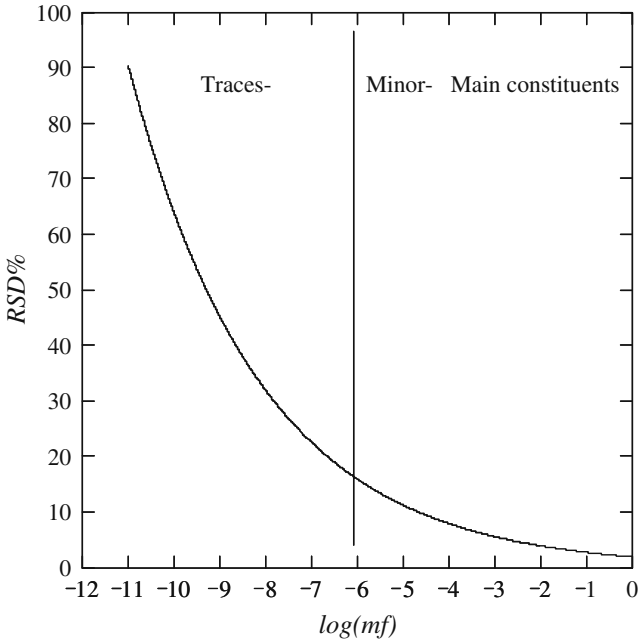

Fig. 2 Graphical illustration of the Horwitz equation. RSD% is the among-laboratory relative standard deviation and *mf* is the analyte mass fraction

reviewing doping cases [18], especially when no credible MU and no proficiency testing data are available, or when no performance requirements are defined [19]. But it must be stressed that the Horwitz equation should never be used as a substitute for the experimental uncertainty estimate of a result.

It was also worthwhile mentioning the need to avoid confusion between the uncertainty associated with the experimental measurements and that associated with setting the conformity limiting values [17]. The latter only reflects uncertainties associated with evaluating the potential deleterious effects of a particular value of the characteristic under examination when the entity subject to conformity assessment is used in a certain context [17]. Situations in which an uncertainty is associated both with the limiting value and with the analytical result were also discussed [7, 10]. But according to the ISO 10576–1 international standard (see the relevant section) the MU should neither explicitly nor implicitly be referred to in the designation of the *LV*s [34].

It was also emphasized that MU cannot be evaluated without metrological traceability. This is particularly mandatory in forensic contexts, in which it is important that MU contains all relevant factors, including all traceability chains [19].

Finally, it was also suggested that, if possible, much more reliable conformity tests can be performed by using uncertainty estimates from interlaboratory comparisons in a learning process [14].

Test for conformity versus tests for non-conformity

A usually neglected aspect when dealing with some limiting value is that, before performing a test, one should decide whether it has to be a test for conformity or a test for non-conformity [12, 13]. In particular, it was emphasized that, if a declaration of conformity with the specifications cannot be stated, it does not mean that the sample under examination is in non-conformity. It can only be stated that the test failed to demonstrate conformity. Likewise, if non-conformity cannot be stated, it does not mean that the sample under examination is in conformity with the specifications [13]. The distinction between the two tests is also considered by the ISO standard (see below) [34] and was already emphasized by Currie [47]: ”Acceptance of a hypothesis, based on statistical testing, must not be taken literally. More correctly, one simply fails to reject the hypothesis in question. For example, non-detection of an analyte does not prove its absence.” and “….Assumption testing, itself, rests upon assumptions. The vast majority of statistical tests performed on the chemical measurement process and its results, for example rely upon the assumptions of randomness and normality”.

So, after a test for conformity, statement A can be claimed: “The measurements have demonstrated, beyond any reasonable doubt, that the value of the measurand is in conformity with the requirements”. On the contrary, after a test for non-conformity, statement B can be claimed: “The measurements have demonstrated, beyond any reasonable doubt, that the value of the measurand is not in conformity with the requirements”. If the result of the selected test is inconclusive, statement C can be claimed “The measurements have not been able to demonstrate, beyond any reasonable doubt, if the value of the measurand is or is not in conformity with the requirements” [13]. These statements perfectly correspond to those reported in paragraphs 7.2–7.4 of the ISO standard [34].

Signal and concentration domains

In general, conformity or non-conformity tests may deal with any type of target variable or measurand. When dealing with most chemical analyses, the measurand is a concentration, however expressed. In this case, tests are relevant to comparison of the concentration of a given analyte in a sample under investigation with a concentration limiting value (or a set of concentration LVs). In these cases, decision making is usually performed in the concentration domain (CD): measurement results with their MU intervals (whatever evaluated) are compared with the proper legal/compositional concentration limiting value, as done in Fig. 1.

When confronting the problem of testing for conformity with a concentration LV, some authors tried to develop alternative approaches enabling decision-making to be performed directly in the signal domain (SD), that is by comparing the signal of the sample under investigation with the signal relevant to a sample containing exactly the specified LV (e.g. to a suitable certified reference material, CRM). Then, the conclusion of the comparison had simply to be translated to the CD by a proper calibration constant. It is well known that the physical quantities of interest (concentration in this case) cannot be measured directly but are connected to the measured signals through a calibration constant [48]. This view is at the basis of some accepted approaches suitable for estimating the limit of detection (LOD), where proper statistical tests must enable evaluation of whether the concentration of the analyte in the sample under investigation is higher than zero (see, for example, among the most authoritative, Refs. [49, 50]). Of course, estimating the LOD is a problem quite similar to that of assessing conformity, where proper statistical test must enable evaluation of whether the concentration of the analyte in the sample under investigation is higher (or lower) than a limiting value.

But comparing signals requires proper consideration of both false-positive (type I) and false-negative (type II) errors: see for example the ISO approach to the LOD [50]. Approaches have been tentatively proposed for assessing conformity to some limiting values by working in the SD and considering both types of errors [4, 6, 8, 11]. Unfortunately, they suffer from disadvantages, namely the actual availability of the CRM containing exactly the concentration of the analyte specified by the limiting value, [6, 8, 11] (quite an improbable case) and, if available, the uncertainty of the analyte concentration in the analysis certificate of the CRM [20]. In subsequent papers, approaches were proposed for performing a test for conformity, or a test for non-conformity, by working in the concentration domain but, at least, by taking into account both false-positive and false-negative errors [20, 23]. In particular, an approach was proposed based on an existing model of the limit of detection [28].

Interpretative problems

Several papers were intended to deal with some peculiar aspects of conformity tests. Attention was focused at cases in which legislative limits were set below the detection capability of the existing analytical techniques [9]. The examined case study showed that legislation based on limiting values may sometimes be beyond analytical capability, and that the limited analytical capability can be exacerbated by the practice of reporting as "undetected" results falling below the limit of detection [9].

The rules for stating when a limiting value is exceeded were introduced in a paper where, after discussing the one-stage and two-stage procedures detailed in Ref. [12] and subsequently adopted by the ISO guidelines [34] (see the section about the ISO guidelines, below), the authors suggested two procedures for estimating the number of measurements necessary for appropriate reliability of the results [12, 13].

Another paper was intended to highlight some experimental problems preventing uniform implementation of legislative standards in the EU and Codex [22]. The authors emphasized that often there is no common interpretation of analytical results across the EU in the food sector, so that significantly different decisions may be taken after analysing the same sample. Particular attention directed at the consequences of reporting and using the experimental results in different ways, considering, or not, the recovery and using results including a different number of significant figures [22].

Economic aspects affecting conformity assessment were also considered by some authors. One paper discussed economic terms of common rules in conformity assessment based on measurement by extending tools of sampling when using *inspection by variable* and *inspection by attribute* [24]. Another paper discussed decision-making in conformity assessment in terms of effective cost associated with measurement, testing and incorrect decision-making [31].

Attention was also drawn to the need to provide an unambiguous and simple procedure for assessing conformity by designing really appropriate decision rules for conformity tests (see the following sections). This should require the knowledge of the acceptable level of the probability of making a wrong decision. The author emphasized that, at least in principle, the acceptable level of the probability of making a wrong decision can be determined if the cost/consequence of taking a wrong decision is known [25]. Unfortunately, as in the case of measurements concerning contaminants in foods, little or no information is usually available [25].

Finally, when introducing the methods used by accredited calibration laboratories, for example within the Deutscher Kalibrierdienst (e.g. the German accreditation body for calibration laboratories, DKD), it was also reported that when measuring a characteristic for conformity with a tolerance zone, e.g. in the case of a set of LVs, a statement of conformity should only be made if the ratio of the width of the tolerance zone to the standard uncertainty associated with the estimate of the characteristic is sufficiently large [27]. Such a condition can be quantified by the measurement capability index, C_m, e.g. the ratio of

the width of permissible values to some multiple of the standard uncertainty associated with the estimate of the characteristic [27] or, analogously, to some multiple of the standard deviation representing the variability of the process or product [32]. This aspect is also managed in the ASME document [33] (see the N:1 decision rule in the relevant section).

International standards

As anticipated in the Introduction, at present the guidelines of ASME [33], ISO [34], and Eurachem/CITAC [35] are among the most useful tools for the operators of testing and calibration laboratories involved in assessing conformity or non-conformity with given specification. This because they describe procedures sufficiently simplified to be widely interpretable and managed.

The ASME document providing guidance for assessment of electrical and mechanical products, is briefly discussed here below, because:

1. its principles inspired the Eurachem/CITAC guide [25, 26, 35]; and
2. it is a convenient introduction to the concept of guard bands and decision rules.

The terminology adopted below can differ from the original terminology used in the three documents because of the need to use the same symbols for the same object/quantities.

The ASME B89.7.3.1-2001 guidelines

The objective of the ASME B89.7.3.1-2001 standard "Guidelines for decision rules: considering measurement uncertainty in determining conformance to specifications" (the ASME guidelines from now on) is to facilitate the development of understanding between suppliers and customers regarding proper consideration of MU in conformity tests [33]. It was prepared by the American Society of Mechanical Engineers, but it can greatly help in understanding the other guidelines because it lists some basic definitions.

According to the ASME guidelines [33], a *decision rule* is a documented rule that describes how MU will be allocated with regard to accepting or rejecting a product according to its specification and the results of a measurement. An *acceptance zone* is the set of values of a characteristic, for a specified measurement process and decision rule, that results in product acceptance when a measurement is within this zone. A *rejection zone* is the set of values of a characteristic, for a specified measurement process and decision rule, that results in product rejection when a measurement is within this zone. A *transition zone* is the set of values of a characteristic, for a specified measurement process and decision rule, that is neither in the acceptance nor rejection zone.

A *guard band* is the magnitude of the offset from the specification limit to the acceptance or rejection zone boundary.

Simple acceptance means a situation when the acceptance zone equals and is identical with the specification zone. *Simple rejection* means a situation when the rejection zone consists of all values of the characteristic outside the specification zone. *Relaxed acceptance* means a situation when the acceptance zone is increased beyond the specification zone by a guard band (Fig. 3a). In a binary decision rule, relaxed acceptance is accompanied by *stringent rejection*. *Stringent acceptance*, in contrast , means a situation when the acceptance zone is reduced from the specification zone by a guard band(s) (Fig. 3b). In a binary decision rule, stringent acceptance is accompanied by *relaxed rejection*.

It should be kept in mind that conformity tests including the choice of the guard band are based on limitations stemming from economic, health, or other fields of interest. The tests performed rely on scientific criteria and limitations, but the final decision is from the outside world.

The ASME guidelines identify different cases of acceptance and rejection zones by decision rules. In particular, it reports that the most common form of acceptance and rejection in industry is that performing simple acceptance and rejection using an N:1 decision rule. N:1 means that the measurement interval, $result \pm U$, cannot be larger than the fraction $1/N$ of the specification zone. Usually N is taken equal to 3 or 4. Of course, using this decision rule can lead to decision making problems when the measurement result is too close to the specification limit (s). Because the N:1 decision rule is not applied by the ISO and Eurachem/CITAC guidelines [34, 35], it is not considered in the following paragraphs.

To increase confidence that a rejected product is actually out of specification, e.g. by choosing a low risk for the producer, ASME guidelines apply a stringent rejection and relaxed acceptance. This means that the relaxed acceptance zone is obtained by increasing the specification zone by a Z% guard band at the specification limit or at both specification limits. Z% is the size of the guard band expressed as a percentage of the expanded uncertainty (a 100% guard band has the magnitude of U) (Fig. 3a).

Similarly, to increase confidence in product quality by reducing the probability of accepting an out-of-specification product, e.g. by choosing a low risk for the consumer, ASME guidelines apply a stringent acceptance and relaxed

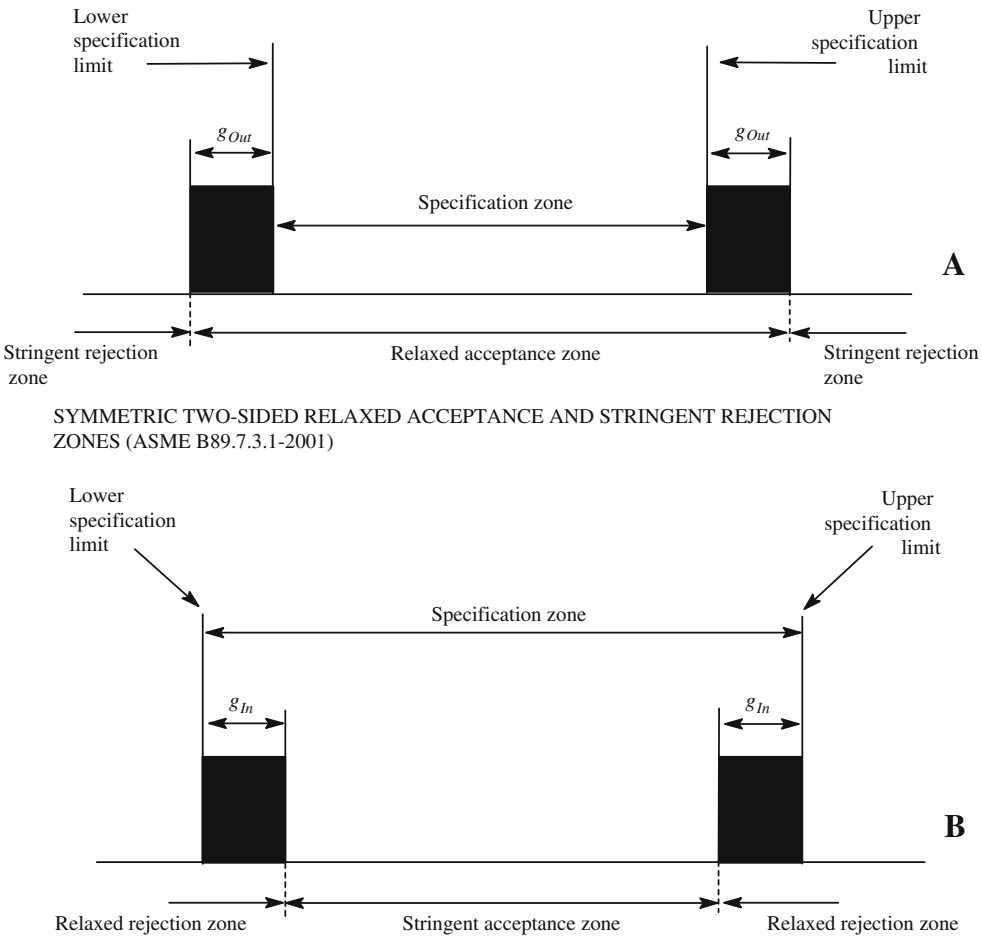

Fig. 3 **a** Symmetric two-sided stringent rejection and relaxed acceptance according to ASME; **b** Symmetric two-sided relaxed rejection and stringent acceptance according to ASME

rejection. Here the acceptance zone is obtained by reducing the specification zone by the guard band(s) amount. Again the size of the guard band is expressed as a percentage of *U* (Fig. 3b).

The ISO 10576–1 international standard

The objective of the ISO 10576–1 international standard (the ISO standard from now on) is to provide assurance of conformity or assurance of non-conformity, either in the form of supplier's declaration, or of a third party certification. In its introduction, the ISO standard, also, provides some definitions. Conformity testing is defined as a systematic examination of the extent to which an entity conforms to a specified criterion [34]. The *limiting values* (*LV*) or *specification limits* (*SL*) are the specified values of the characteristic giving upper and/or lower bounds of the permissible values. The *region of permissible values* is the interval or intervals of all permissible values of the characteristic. The *region of non-permissible values* is the interval or intervals of all values of the characteristic that are not permissible (Fig. 4). The intervals are based on accepted and required probabilities.

The ISO standard also details the requirements for defining limiting values. The entity and the quantifiable characteristic of the entity shall be clearly and unambiguously specified, the test procedure should be a standardised one and, as already cited in the section *Which uncertainty?*, the MU shall neither explicitly nor implicitly be referred to in the designation of the *LV*s. The ISO standard reports examples of single and double *LV*s and specifies that the uncertainty interval shall be determined according to the GUM [1].

The principal feature of the ISO standard is the recommendation that the conformity test be performed as a two-stage procedure, in agreement with a previous suggestion [12, 13]. The advantage of the two-stage procedure is a substantially higher probability of declaring conformity for entities with permissible values of the quantity of interest (the concentration, in chemical analyses) which are closer to the *LV*. The two-stage procedure is represented in Fig. 5. By the wording "appropriate combination of the two (set of) measurement results"

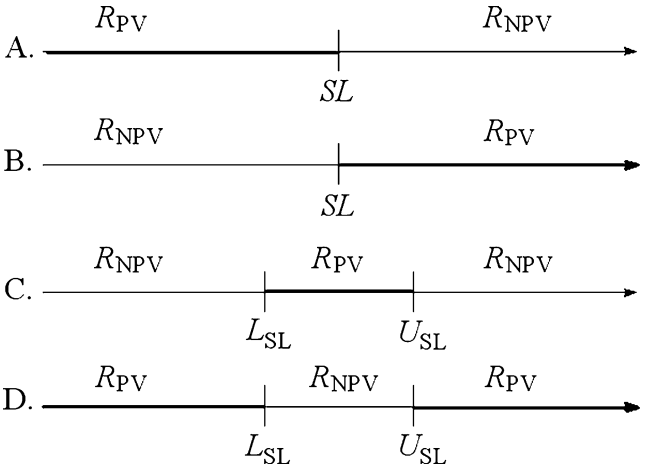

Fig. 4 Division of the domain of the characteristic in regions of permissible and non-permissible values according to ISO 10576–1. R_{PV}, region of permissible values; R_{NPV}, region of non-permissible values. **a** Case of an upper *LV*; **b** case of a lower *LV*; **c** First case of double limits (the region of permissible values is within the limits); **d** second case of double limits (the region of permissible values is outside the limits)

(Fig. 5) it is meant that, in the second stage, the decision can be taken by computing the average estimate of the quantity of interest and its uncertainty interval by using the n_2 results obtained in the second stage only or those obtained in the second stage plus the n_1 results obtained in the first stage. The one-stage procedure is applied when the two-stage procedure is not necessary, because the first stage enables the necessary decision making, or when it cannot be performed by some reason. Of course, the one-stage procedure stops at the end of the first stage of Fig. 5. Conformity/non-conformity may be assured if, after performing the measurement procedure and calculating the MU, the estimated uncertainty interval of the measurement result is inside the region of permissible/non-permissible values.

The ISO standard introduces both tests for conformity and for non-conformity, by specifying the following possibilities of reporting the results of the conformity assessment:

- Assurance of conformity: *the conformity test has demonstrated beyond any reasonable doubt that the value of the characteristic is in conformity with the requirements.*
- Assurance of non-conformity: *the conformity test has demonstrated beyond any reasonable doubt that the value of the characteristic is not in conformity with the requirements.*
- Inconclusive result: *the conformity test has not been able to demonstrate beyond any reasonable doubt that the value of the characteristic is or is not in conformity with the requirements.*

The Annex B of the ISO standard reports few illustrative examples [34].

The Eurachem/CITAC guide

The Eurachem/CITAC guide (the Guide from now on) [35] was issued more recently than the ASME and ISO standards and, as already reported, it follows the principles outlined in ASME B89.7.3.1-2001 [33]. The principles of the Guide were also detailed in a recent paper [26]. As already reported, the Guide describes typical scenarios arising when some measurement result is used for assessing compliance with an upper *LV* according to Fig. 1. As with the ASME guidelines [33] and ISO standards [34], the Guide assumes that uncertainty has been evaluated by the method provided by ISO [1] and Eurachem [43] and includes the uncertainty of sampling. Most of definitions are equivalent to those given by ASME and ISO. As done by ASME, the Guide emphasizes that the key to the assessment of conformity is the concept of decision rules. Decision rules enable determination of Acceptance and Rejection zones. The zones are determined in such a way that if the measurement result lies in the acceptance zone the product is in conformity with the requirements while, if it lies in the rejection zone, it is in non-conformity with the specification. In mentioning the different zones, the Guide does not mention simple, stringent, and relaxed zones as ASME does.

In addition, the Guide presents cases of more or less simple decision rules. In particular, it gives details of a decision rule set up by the *Article* 6 - *Interpretation of results* of Directive 96/23/EC [36] (see the next section):

1. The result of an analysis shall be considered non-compliant if the decision limit of the confirmatory method for the analyte is exceeded.
2. If a permitted limit has been established for a substance, the decision limit is the concentration above which it can be decided with a statistical certainty of $1-\alpha$ that the permitted limit has been truly exceeded.
3. If no permitted limit has been established for a substance, the decision limit is the lowest concentration level at which a method can discriminate with a statistical certainty of $1-\alpha$ that the particular analyte is present.
4. For substances listed in Group A of Annex I to Directive 96/23/EC, the α error shall be 1% or lower. For all other substances, the α error shall be 5% or lower.

As emphasized by the Guide, such statements correspond to a decision of non-conformity or rejection with low probability of false rejection (high confidence of correct rejection) (Fig. 6a). It is easily observed that, in practice,

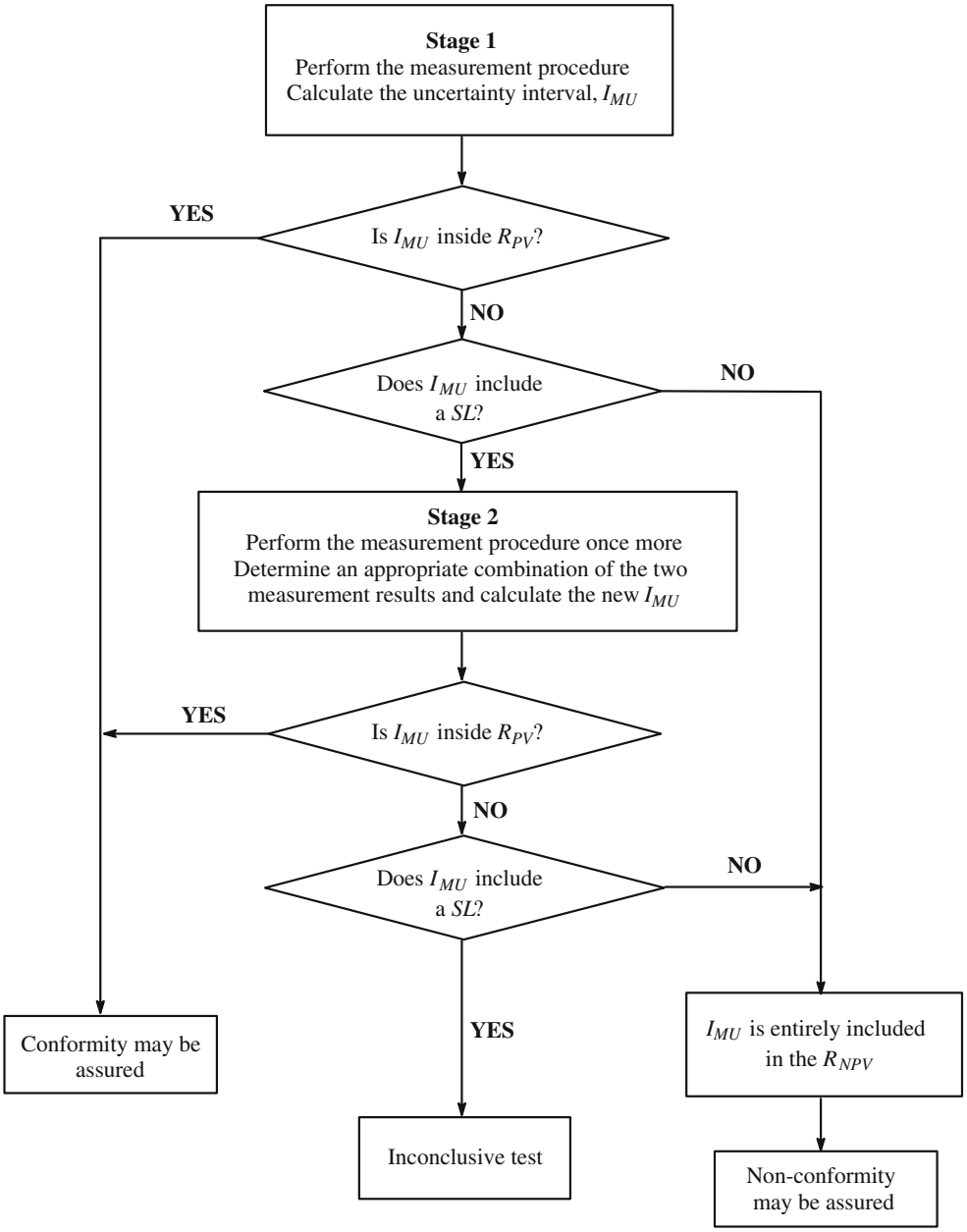

Fig. 5 Flow diagram for the two-stage procedure. I_{MU}: uncertainty interval

the acceptance zone in Fig. 5a corresponds to the relaxed acceptance zone according to ASME (reported in Fig. 3a). In the case presented above, the value of the guard band, g, is chosen so that, for a measurement result greater than or equal to $LV+g$, the probability of false rejection is less than or equal to α. A typical value of α (the probability of false positive errors) is 5%.

Analogously, in Fig. 6b the acceptance zone corresponds to the stringent acceptance zone according to ASME guidelines (reported in Fig. 3b). Some potentially unclear aspects of the Eurachem/CITAC definition of acceptance and rejection zones were recently discussed [29, 30].

Appendix A of the Guide reports some examples of how the guard bands can be determined. In general the size of the guard band is $k \cdot u$ (see the section *Which uncertainty*?). A point deserving some comment is relevant to Case 1a, in which only the standard uncertainty, u, is available [35]. It is reported that in many cases, current practice is to use $k=2$. As stated by the Guide, on the assumption that the distribution is approximately normal, this choice gives a level of confidence of approximately 95% that the value of the measurand lies in the interval $y \pm 2 \cdot u$. On this basis, the Guide states that "the probability that the value of the measurand is less than $y+2 \cdot u$ is approximately 97.5%." It is likely that this last sentence can be quite perplexing to

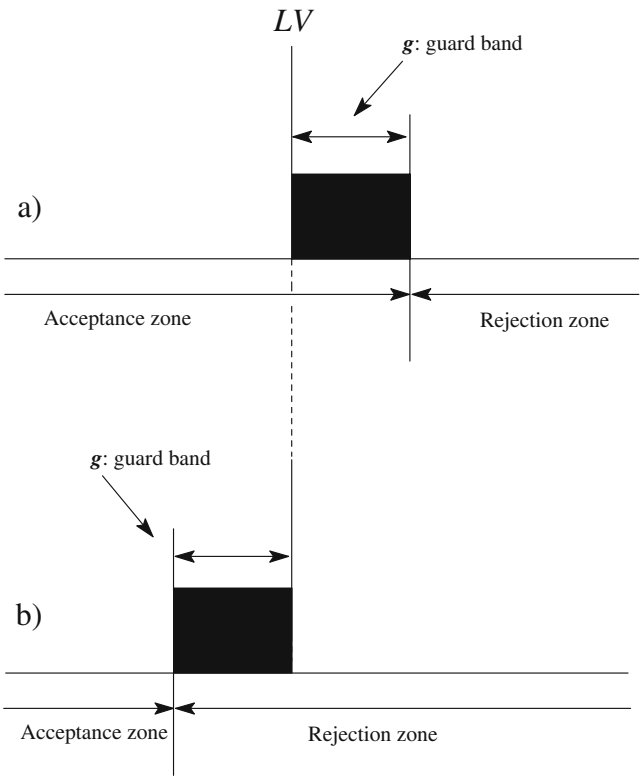

Fig. 6 Acceptance and rejection zones for an upper LV according to Eurachem/CITAC

readers not well trained in probability distributions. Figure 7a enables elucidation of the difference between 95%, e.g. the percentage area of the distribution included in the $y \pm 2u_c$ interval, and 97.5%, e.g. the area at the left of the $y+2u_c$ value. Figure 7b displays the situation for a guard band equal to one u_c. The reported values are those obtained by the following equations for Fig. 7a:

$$\int_{y-2u_c}^{y+2u_c} f(c)dc \cong 0.954 \tag{2}$$

$$\int_{0}^{y-2u_c} f(c)dc \cong 0.023 \cong \int_{y+2u_c}^{\infty} f(c)dc \tag{3}$$

and by the following equations for Fig. 7b:

$$\int_{y-u_c}^{y+u_c} f(c)dc \cong 0.683 \tag{4}$$

$$\int_{0}^{y-u_c} f(c)dc \cong 0.159 \cong \int_{y+u_c}^{\infty} f(c)dc \tag{5}$$

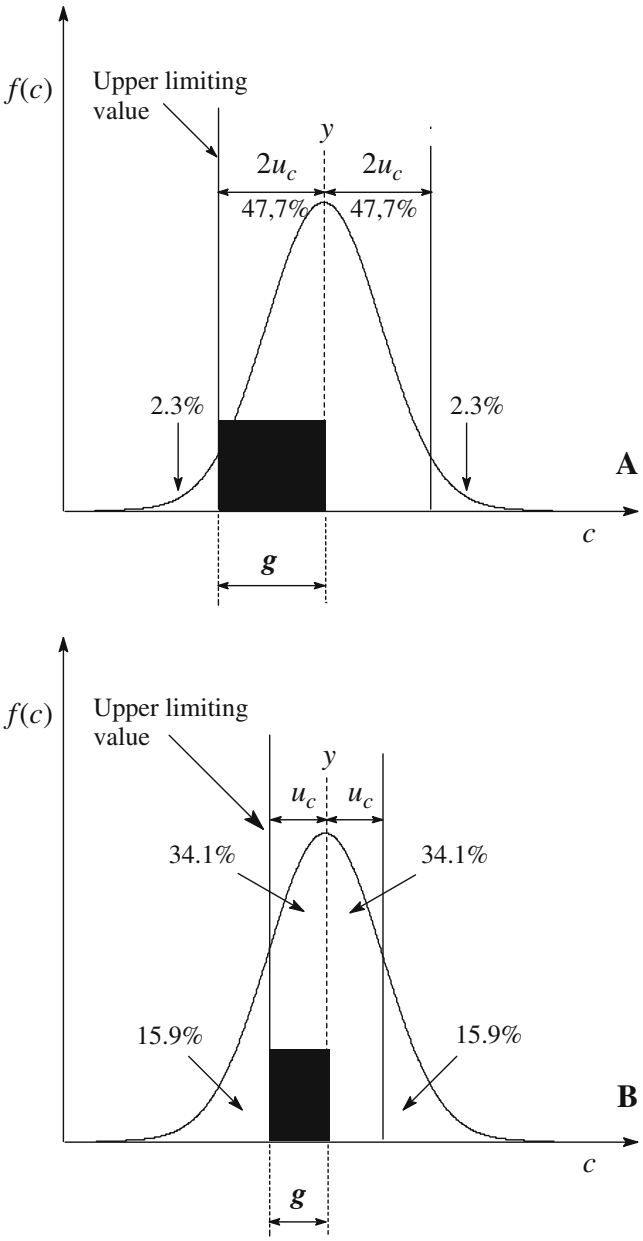

Fig. 7 **a** Explanation of the sentence "If the size of the guard band is $2{\cdot}u$, then the probability that y is less than $y+2{\cdot}u$ is approximately 97.5%". **b** same as **a** but the size of guard band is u

About Directive 96/23/EC

The Commission Decision of 12 August 2002 (the Decision from now on) was aimed at implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results obtained in the monitoring of specific substances and residues thereof in live animals and animal products, when they affect public health (available online) [36]. The Decision provides rules for the analytical methods to be used in the testing of official samples and specifies common criteria for the

interpretation of analytical results of official control laboratories for such samples. The *Article* 6 - *Interpretation of results* is reported here in the section dealing with the Eurachem/CITAC guide.

In the Annex "Performance Criteria, Other Requirements and Procedures for analytical methods" the Decision gives the following definitions:

1.11. Decision limit (CCα) means the limit at and above which it can be concluded with an error probability of α that a sample is non-compliant.

1.12. Detection capability (CCβ) means the smallest content of the substance that may be detected, identified and/or quantified in a sample with an error probability of β. In the case of substances for which no permitted limit has been established, the detection capability is the lowest concentration at which a method is able to detect truly contaminated samples with a statistical certainty of $1-\beta$. In the case of substances with an established permitted limit, this means that the detection capability is the concentration at which the method is able to detect permitted limit concentrations with a statistical certainty of $1-\beta$.

CCα and CCβ are concentration values. It follows that CCα is the upper limit of the region of permissible concentration values. The Decision should represent a qualified reference for operators of laboratories accredited for official residues control. Unfortunately, it was shown that some statements of the Decision can generate misunderstanding and/or confusion [21]. For example, Articles 3.1.2.5 and 3.1.2.6 of the Decision recommend estimation of CCα and CCβ according to ISO 11843 [51]. But ISO 11843 uses different symbols (x_C and x_D in place of CCα and CCβ), considers sample statistics (*t*-distributions) in place of population statistics (normal distributions) and estimates the two limits by use of somewhat different approaches [21]. Moreover, the Annex of the Decision explains the meaning of CCα in the Fig. 3.2 of the Decision. But that figure can mislead the reader, because it reports CCα in a frequency versus *response* diagram [21] whereas CCα is a concentration. In contrast, the subsequent Fig. 3 of the Decision correctly shows CCβ in a frequency versus *concentration* diagram.

Worked examples

Worked examples in such a tricky matter as conformity testing are hardly representative of the plethora of possible experimental situations. Nevertheless, some examples representative of basic experimental situations are presented here to help readers evaluate ISO and Eurachem/CITAC approaches.

Example 1

The first example deals with the ISO approach [34] according to the flow diagram reported in Fig. 5 (the two-stage procedure).

The 98/83/EC directive on the quality of water intended for human consumption specifies the upper limit value LV_u=10 μg L^{-1} for the concentration of arsenic in drinking water [52]. When using a two-stage procedure, the sample is divided into two subsamples, and the second is only used if the uncertainty interval, I_{MU}, contains the limiting value. The ISO approach accepts uncertainty intervals given in the form of a confidence interval (subclause 6.4). Suppose that the measurements are performed with a standard measurement procedure which operates with a combined standard uncertainty of u_c= 1.485 μg L^{-1} at concentration levels around the LV_u. According to the ISO approach (Annex B, Example 2), if n independent measurements, each with uncertainty σ_Y, are performed and the arithmetic mean of the measurements is $\overline{Y}$, then the confidence interval is given as:

$$\overline{Y} \pm \frac{z_{1-\frac{\alpha}{2}}, \sigma_Y}{\sqrt{n}} \tag{6}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

A first series of three independent analyses of the arsenic concentration in the first water subsample gives the concentration $\overline{C}_{As,1} = 9.09 \mu gL^{-1}$. Using $z_{1-\frac{\alpha}{2}} = 1.96$ (often approximated to 2.0) to choose α=0.05, one can obtain the uncertainty interval:

$$I_{MU} = 9.09 \pm \frac{1.96 \cdot 1.485}{\sqrt{3}} = 9.09 \pm 1.68\ \mu gL^{-1} \tag{7}$$

Because the upper limit value, LV_u=10 μg L^{-1}, is within the uncertainty interval, the test is inconclusive at the given confidence level.

A second series of four independent analyses of the arsenic concentration is then performed with the second water subsample. This gives the result $\overline{C}_{As,2} = 8.66 \mu gL^{-1}$. The uncertainty interval is now:

$$I_{MU} = 8.66 \pm \frac{1.96 \cdot 1.485}{\sqrt{4}} = 8.66 \pm 1.46\ \mu gL^{-1} \tag{8}$$

Again the upper limit value, LV_u=10 μg L^{-1}, is within the uncertainty interval and the test is inconclusive.

The results of both set of measurements are the combined. The concentration of arsenic resulting from the seven measurements is:$\overline{C}_{As} = \frac{\overline{C}_{As,1} \cdot 3 + \overline{C}_{As,2} \cdot 4}{7} = 8.84\ \mu gL^{-1}$. The new I_{MU} is:

$$I_{MU} = 8.84 \pm \frac{1.96 \cdot 1.485}{\sqrt{7}} = 8.84 \pm 1.10\ \mu gL^{-1} \tag{9}$$

This uncertainty interval is all below the LV_u. This last result enables the statement of conformity with the specified limit at the given confidence level.

Example 2

The second example deals with the Eurachem/CITAC approach [35]. In this case, one should define the size of the guard band and choose to perform a test enabling high confidence of correct rejection (as in Fig. 6a) or high confidence of correct acceptance (as in Fig. 6b).

Commission regulation (EU) No 105/2010 of 5 February 2010 amending Regulation (EC) No 1881/2006 setting maximum levels for specific contaminants in foodstuffs with regard to ochratoxin A specifies the upper limit value LV_u=80 μg kg^{-1} for the concentration of ochratoxin A in liquorice extract [53]. Suppose that the combined uncertainty of a measurement method, evaluated before performing the necessary measurements, is 3.5 μg kg^{-1}. By using the above data, choosing α=0.05 and performing the test enabling high confidence of correct acceptance, the guard band is equal to g=1.65·3.5=5.775 μg kg^{-1}. The guard band is subtracted from the upper limit value. The acceptance zone then extends to 80–5.775= 74.225 μg kg^{-1}. Suppose that the concentration of ochratoxin A in two samples are 86.07 μg kg^{-1} and 72.33 μg kg^{-1}. The first sample is rejected and the second is accepted. In contrast, when performing the test is to enable high confidence of correct rejection, the guard band is added to the upper limit value. The acceptance zone then extends to 80+5.775=85.775 μg kg^{-1}. In this second case, both samples are accepted.

Conclusions

This literature information confirms that some aspects of the assessment of conformity with legal or compositional limiting values deserve further developments. A general agreement exists about the need to properly take into account the uncertainty of measurement in decision making, and use of the correct MU (including the sampling uncertainty component and estimated according to GUM). But problems still exist, especially concerning the need for unification and/or unequivocal formulation of the wording of prescriptions by the regulatory Authorities. Finally, decision making could be greatly facilitated by issuing really unified and, consequently, generally agreeable and usable guidelines.

Acknowledgements Financial support by the Fondo Interno per la Ricerca Scientifica e Tecnologica 2008 (Università degli Studi di Milano) is acknowledged.

References

1. Guide to the expression of Uncertainty in Measurement (1995) 1st Ed., and corrected and reprinted Ed., ISO Secretariat, Switzerland
2. Analytical Methods Committee (1995) Uncertainty of measurement: implications of its use in analytical science. Analyst 120:2303–2308
3. Barwick V (1996) Setting and enforcing legislative limits. Anal Commun 33:123–124
4. de Ruig WG (1996) Characterizing and optimizing analytical method quality. Trends Food Sci Technol 7:234–240
5. Ellison S, Wegscheider W, Williams A (1997) Measurement uncertainty. Anal Chem 69:607A–613A
6. Desimoni E, Mannino S (1998) Uncertainty of measurement and legislative limits. Accredit Qual Assur 3:335–337
7. King B (1999) Assessment of compliance of analytical results with regulatory or specification limits. Accredit Qual Assur 4:27–30
8. Buzoianu M (2000) Measurement uncertainty and legal limits in analytical measurements. OIML Bull XLI:5–12
9. Ellison SRL, Barwick VJ, Williams A (2000) Legislative limits below detection capability. Accredit Qual Assur 5:308–313
10. Hibbert DB (2001) Compliance of analytical results with regulatory or specification limits: a probabilistic approach. Accredit Qual Assur 6:346–351
11. Desimoni E, Mannino S, Brunetti B (2001) On the assessment of compliance with legal limits: Part 1: signal and concentration domains. Accredit Qual Assur 6:452–458
12. Holst E, Thyregod P, Wilrich P-Th (2001) On conformity testing and the use of two stage procedures. Int Stat Rev 69:419–432
13. Christensen JM, Holst E, Olsen E, Wilrich PT (2002) Rules for stating when a limiting value is exceeded. Accredit Qual Assur 7:28–35
14. Andersson H (2002) The use of uncertainty estimates of test results in comparison with acceptance limits. Accredit Qual Assur 7:228–233
15. Glaeser H (2002) Compliance with legal requirements: definition of limits, sampling and measurement uncertainty. Accredit Qual Assur 7:388–391
16. Kessel W (2002) Measurement uncertainty according to ISO/BIPM-GUM. Thermochim Acta 382:1–16
17. Källgren H, Lauwaars M, Magnusson B, Pendrill L, Taylor P (2003) Role of measurement uncertainty in conformity assessment in legal metrology and trade. Accredit Qual Assur 8:541–547
18. van der Veen AMH (2003) Measurement uncertainty and doping control in sport. Accredit Qual Assur 8:334–339
19. van der Veen AMH (2004) Measurement uncertainty and doping control in sport. Part 2: Metrological traceability aspects. Accredit Qual Assur 9:311–316
20. Desimoni E, Brunetti B, Cattaneo R (2004) The assessment of compliance with legal limits. Part 2. How to consider both type I and type II errors when working in the concentration domain. Accredit Qual Assur 9:445–449
21. Desimoni E, Brunetti B (2004) About CCα and CCβ as introduced by the commission decision of 12 August 2002 implementing council directive 96/23/EC. Accredit Qual Assur 9:724–725
22. Wood R (2006) Compliance decision criteria—problems encountered and actions taken in the EU and Codex. Accredit Qual Assur 11:83–88
23. Desimoni E, Brunetti B (2006) Considering uncertainty of measurement when assessing compliance or non-compliance with reference values given in compositional specifications and statutory limits: a proposal. Accredit Qual Assur 11:363–366

24. Pendrill LR (2008) Operating 'cost' characteristics in sampling by variable and attribute. Accredit Qual Assur 13:619–631
25. Williams A (2008) Compliance with specifications. Accredit Qual Assur 13:617–618
26. Williams A (2008) Principles of the EURACHEM/CITAC guide "Use of uncertainty information in compliance assessment". Accredit Qual Assur 13:633–638
27. Czaske M (2008) Use of uncertainty measurement by accredited calibration laboratories when stating compliance. Accredit Qual Assur 13:645–651
28. Desimoni E, Brunetti B (2008) About considering both false negative and false positive errors when assessing compliance and non-compliance with reference values given in compositional specifications and statutory limits. Accredit Qual Assur 13:653–662
29. Desimoni E, Brunetti B (2010) About Acceptance and Rejection zones as defined in the EURACHEM/CITAC Guide (2007) "Use of uncertainty information in compliance assessment". Accredit Qual Assur 15:45–47
30. Ellison SLR, Williams A (2010) Response to "About acceptance and rejection zones". Accredit Qual Assur 15:49–51
31. Beges G, Drnovsek J, Pendrill LR (2010) Optimising calibration and measurement capabilities in terms of economics in conformity assessment. Accredit Qual Assur 15:147–154
32. Hinrichs W (2010) The impact of measurement uncertainty on the producer's and user's risks, on classification and conformity assessment: An example based on tests on some construction products. Accredit Qual Assur 15:289–296
33. ASME B89.7.3.1-2001 (2002) Guidelines for decision rules: considering measurement uncertainty in determining conformance to specifications, ASME, New York
34. ISO 10576–1:2003 (2003) Statistical methods – Guidelines for the evaluation of conformity with specified requirements, Part 1: General principles, ISO, Geneva
35. Ellison SRL, Williams A (eds.) (2007) EURACHEM/CITAC Guide "Use of uncertainty information in compliance assessment" first edn 2007. http://www.eurachem.org/. Accessed 17 September 2010
36. Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results (2002/657/EC). http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!DocNumber&lg=en&type_doc=Decision&an_doc=2002&nu_doc=657. Accessed 17 September 2010
37. ILAC-G8 (1996) Guidelines on assessment and reporting of conformity with specification, International Laboratory Accreditation Cooperation, ILAC, Australia
38. NMKL Procedure No. 5 (1997) Estimation and expression of measurement analysis in chemical analysis, NMKL, Finland
39. UKAS LAB 12 (2000) The expression of uncertainty in testing, UKAS, UK
40. Wunderli S (2003) Uncertainty and sampling. Accredit Qual Assur 8:90
41. Heydorn K, Esbensen K (2004) Sampling and metrology. Accredit Qual Assur 9:391–396
42. Ramsey MH, Thompson M (2007) Uncertainty from sampling, in the context of fitness for purpose. Accredit Qual Assur 12:503–513
43. Ellison SRL, Rosslein M, Williams A (eds.) (2000) EURACHEM/CITAC Guide "Quantifying uncertainty in analytical measurements" QUAM: 2000.P1. http://www.eurachem.org/. Accessed 17 September 2010
44. Horwitz W, Kamps LR, Boyer KW (1980) J Assoc Off Anal Chem 63:1344–1354
45. Horwitz W (1982) Evaluation of analytical methods used for regulation of foods and drugs. Anal Chem 54:67A–76A
46. Albert R, Horwitz W (1997) A Heuristic derivation of the horwitz curve. Anal Chem 69:789–790
47. Linsinger TPJ, Josephs RD (2006) Limitations of the application of the Horwitz equation. Trends Anal Chem 25:1125–1130
48. Currie LA (1995) Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). Pure Appl Chem 67:1699–1723
49. Currie LA (1968) Limits for qualitative detection and quantitative determination. Application to radiochemistry. Anal Chem 40:586–593
50. International Organization for Standardization (ISO) ISO (1997) 11843-1 Capability of detection – Part 1:L Terms and definitions
51. ISO 11843: 1997 (1997) Capability of detection-part 1: terms and definitions, part 2: methodology in the linear calibration case, ISO, Geneva
52. Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption, OJ C, 5.12.98, p. 32
53. Commission regulation (EU) No 105/2010 of 5 February 2010 amending Regulation (EC) No 1881/2006 setting maximum levels for certain contaminants in foodstuffs as regard ochratoxin A OJ L, 6.2.2010, p. 7