Recommendations to make reliable the determinations of testing uncertainties

Louis-Jean Hollebecq
Scientific and technical Manager

Table of content

Annex:

Examples of implementing different methods for the determination of uncertainties.

Abstract:

Results of CompaLab ILC (interlaboratory comparisons) show that uncertainties are significantly underestimated by participants. A gradation of test methods can be established, from mainly metrological to methods which sources of uncertainties are mainly qualitative. Uncertainties are globally well determined for the first while they are globally underestimated by a factor 10 or more for the last. This probably comes from a massive choice of GUM method B to determine them, whatever the test method. However, method B is effective in metrology but not when significant qualitative sources of uncertainty are present. GUM also lacks guidance about some issues specific to testing. Furthermore, ILC and laboratory quality surveillance results can be re-used for GUM method A, which provide quite better estimates of uncertainties and request significantly fewer time and money than method B. When accurate determination of uncertainties is important, collaborative method A experiments (i.e. specifically designed ILC) should be organised, which results can afterwards be used in very effective internal quality surveillance programs. Determining uncertainties should always begin by a clarification about the intended use of them and a collection of available information concerning the precision of testing. The most appropriate method to determine uncertainties highly depends on this and, in most cases, the answer is not method B.

1 Introduction

Since 2005 and confirmed in 2017, laboratories accredited against ISO/IEC 17025 [1] have to determine the uncertainties that are linked to their test results. To achieve it, they mostly use the methods (particularly the method B) of GUM [2] which was issued by BIPM. These methods are based on principles detailed in its annex E and summarised here after:

- The uncertainties should be estimated in the most possible realistic way, contrarily to former usual habits that were to determine uncertainties by excess to make sure that the reference value is inside the confidence interval;
- None of the errors are completely random (i.e. with mean value = 0) or completely systematic (i.e. with standard deviation = 0). For this reason, they should be all treated in the same way.

Consequently:

- Uncertainties should all be treated as standard deviations;
- When the bias is known, the test result shall be corrected from it;
- When the bias is unknown, it shall be treated as if it were a random component to uncertainty.

This document is very valuable because it is based on a very long experience of determining uncertainties in the field of metrology. Consequently, it is a very detailed document addressing many difficult issues raised when a calculation of uncertainties is needed.

However, as it was produced by metrologists for metrologists, some major features specific to laboratory testing are not well dealt with in GUM, particularly the bias. On the other hand, ISO 13528 [3] provides tools (ζ scores) for assessing the uncertainty claimed by a laboratory independently to their bias during PT. CompaLab performance of this assessment showed that laboratories massively underestimate their uncertainties.

To deal with this issue, this document:

- Provide results of assessment of uncertainties found in CompaLab ILC during the last years;
- Discuss the basics of uncertainty in the specific context of lab testing;

- Explores the differences of conditions between metrology and laboratory testing that may lead to differences in the methods with which uncertainties should be determined;
- Establish a list of possible methods to determine test uncertainties and of their advantages and disadvantages;
- Discuss some practical issues that may improve the quality of determinations of uncertainties;
- Provide some proposals for labs to assess the quality of the determination of their uncertainties.

The aim of this document is not to duplicate or to explain the content of the reference documents, in particular GUM [2] and VIM [4], but to provide guidance on when and how to use them or other technics to achieve an efficient determination of uncertainties.

2 Symbols and abbreviations

The symbols used in this document are listed in Table 1.

Table 1. List of symbols used in this document.

Symbol	**Designation and comments**
B	Bias of a test result
c	Weighting coefficient
CoV	Coefficient of variation, defined by $CoV = \mu/\sigma$
e	Random error on a test result
k	Enlargement coefficient of uncertainties
i	Rank of an item or value
IC	Interval of confidence with enlargement coefficient k taken equal to 2
m	Estimate of a mean value
Med	True median value
n	Total number of elements of a series
N	Total number of random values
r	Correlation coefficient
s	Estimate of a standard deviation
s_i	i^{th} estimate of an ordered series of estimates of standard deviations
s_L	Estimate of interlaboratory standard deviation
s_r	Estimate of repeatability standard deviation
s_R	Estimate of reproducibility standard deviation
u	Standard uncertainty
u_{Xpt}	Standard uncertainty on an assigned value in an ILC
u_{RM}	Standard uncertainty on an assigned value of a reference material
U	Enlarged uncertainty, with enlargement coefficient “2”
$\overline{x}$	Estimate of a mean value of a series of *x*
X_{pt} ***and*** X_{ILC}	Assigned value in an ILC

Symbol	**Designation and comments**
X_{Lab}	Mean value of the test results of a laboratory
X_{RM}	Assigned value of a reference material
μ	True value of a mean value
σ	True value of a standard deviation
χ^2_{n-1}	Value of the Khi² distribution law with n-1 degrees of freedom
ζ	ζ-score for assessing uncertainties as described in ISO 13528 [3]

Abbreviations:

- CRM: certified reference material;
- IC: interval of confidence;
- ILC: interlaboratory comparison;
- PT: proficiency testing;
- RM: reference material;
- SD: standard deviation.

3 Technical backgrounds

3.1 Basics of ISO 13528 for assessing the uncertainties

ISO 13528 [3] uses so-called ζ-scores as defined in Equation (1) to assess uncertainties claimed by laboratories participating to PT programs.

$$\zeta = \frac{x_i - X_{pt}}{\sqrt{u_{Xpt}^2 + u_i^2}} \quad (1)$$

where x_i *is the result of participant* i*,*
X_{pt} *is the assigned value for the ILC,*
u_{Xpt} *is the uncertainty on the assigned value,*
and u_i *is the uncertainty claimed by the participant for its result.*

The basics for this Equation (1) is to compare:

- The gap between the result of the participant and a value regarded as a reference value;
- And the combination of the uncertainty it claims on its results and on the reference value.

If the uncertainty claimed by the lab is too small, the gap between its results and the reference value becomes incompatible with this claimed uncertainty and the ζ-score overcomes limits that are usually set up to 2 and 3, in reference to usual limits used for normal distributions.

It shall be noted that this ζ-score can only detect uncertainties that are obviously underestimated. When the uncertainty is over estimated, the corresponding ζ-score tends to 0, but a situation where ζ-score tends to 0 car also mean that the result of the participant is by chance very close to the reference value. In CompaLab, we then decided to signal the situations where the uncertainties are likely to be overestimated by comparing them to the

reproducibility SD: a warning is triggered when $u_i > 5.s_R$. This is based on statements that are provided further on in § 11. It should however be immediately noted that coefficient 5 is completely conventional, and it would have made sense to choose a lower coefficient.

3.2 Results of assessments performed within CompaLab during the last years

Typical results of assessment of uncertainties are represented in Figure 1.a to c.

In abscissa of these figures, participants are ranked by increasing ζ-score. The scale for ordinates is graduated in units of s_R. For example, in Figure 1.a, s_R = 0,0062% mass. Then, the gap between ordinate +5 and ordinate -5 is 0,216-0,154 = 0,062% corresponding to $10.s_R$. The line "0" corresponds to the X_{pt} value and the dashed lines represent the limit for u_{Xpt} on each side of the X_{pt} line.

Dots represent the participants results: in green for those with $\zeta < 2$, orange for those with $2 < \zeta < 3$, and red for those with $\zeta > 3$. In addition to that, yellow dots circled with green represent participants for which $\zeta < 0,5$ and $u_i > 5.s_R$. The uncertainty claimed by these participants is likely to be overestimated. Vertical segments are representing the u_i value of each of the participants: the more the limits of segments are away from the the X_{pt} line, the larger the ζ-score is.

It shall be reminded that, in CompaLab PT programs, participants are not obliged to provide figures for uncertainty. It follows that the figures provided are those that participants had determined prior to their participation, and are not an exercise that they performed only to satisfy the requirements of the PT program.

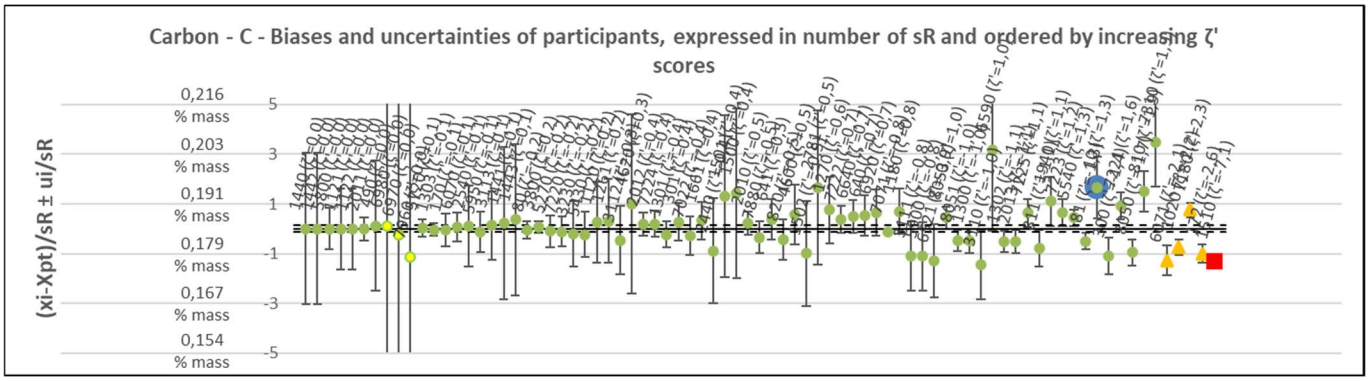


Figure 1.a: Results of 2023 of assessment of uncertainties for carbon content on low alloyed steel.

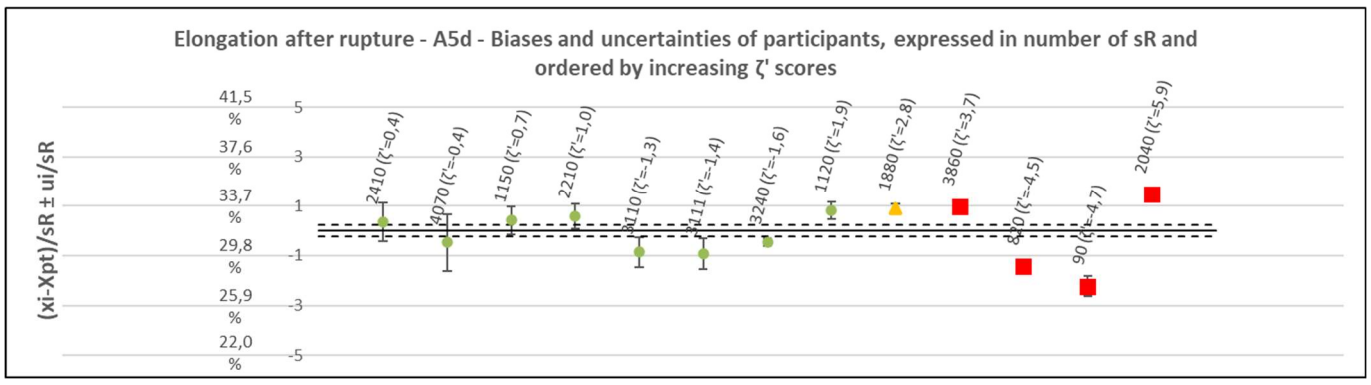


Figure 1.b: Results of 2023 of assessment of uncertainties for A_{5d} (elongation after rupture) in a tensile test on a carbon steel.

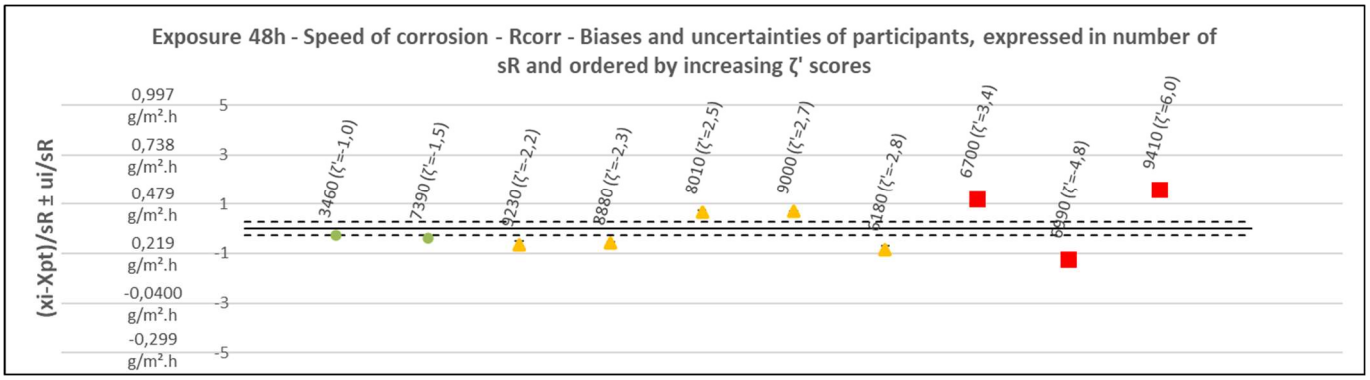


Figure 1.c: Results of 2023 of assessment of uncertainties for speed of intergranular corrosion in a Huey test on a stainless steel.

The 3 examples of Figure 1 represent 3 typical types of situations:

- Case 1, illustrated by Figure 1.a: test results that are mainly governed by metrological issues. In addition to chemical content of major elements, it can be hardness test results or even mechanical properties related to forces like tensile strength. In this case, most of the lengths of vertical segments are in the range of s_R and few participants fail to provide relevant uncertainties. Moreover, the uncertainty figures of poor quality are balanced between the underestimated and the overestimated ones;
- Case 2: illustrated by Figure 1.b: test results that are in between case 1 and case 2, for which both metrological and technological issues have got some importance. In this case, most of the lengths of vertical segments are smaller than s_R and an abnormal proportion of participants provided underestimated uncertainty figures;
- Case 3: illustrated by Figure 1.b: test results that are mainly governed by technological issues. In this case, the length of vertical segments is always lower than the gap between the X_{pt} and the dashed lines, i.e. all u_i are lower than u_{Xpt} although we have every reason to believe the opposite. Almost all participants got an alert and those who do not, probably do not by chance.

3.3 Basics for uncertainty that are common to metrology and testing

The definition of uncertainty is similar in metrology and in testing. In both cases, the definitions of VIM [4] apply, among which some (in particular repeatability and reproducibility) come from ISO 3534-2 [5] via ISO 5725-1 [6]. Uncertainty is defined as a non-negative parameter that provides information about the maximum distance that is likely to exist between a measured quantity value and a supposed reference value.

A model for evaluating that distance is provided by ISO 5725-1 [6] in the form of Equation (2) as follows:

$$y = m + B + e \qquad (2)$$

where y *is the measured quantity value (test result in the case of lab testing),*
m *is the reference value,*
B *is the laboratory component of bias in repeatability conditions,*
e *is the random error occurring in any measurement in repeatability conditions.*

ISO 5725-1 [6] also introduces the notions of:

- Accuracy, defined as the closeness of agreement between a test result and the true value;
- Trueness, defined as the closeness of agreement between the expectation of test result and the true value, usually expressed as a bias;

- Precision, defined as the closeness of agreement between independent test results obtained in stipulated conditions, usually expressed as a standard deviation;
- Repeatability, defined as the precision in conditions where independent test results were obtained with the same method, on identical test or measurement items, in the same test or measurement facility, by the same operator, using the same test equipment, within short intervals of time;
- Reproducibility, defined as the precision in conditions where independent test results were obtained with the same method, on identical test or measurement items, in different test or measurement facilities, by different operators, using different test equipment;

Comments concerning *m*:

1. To be implemented, Equation (2) requests the definition of the *m* value. It is now widely accepted that no "true value" ever exists, at least for test results (but also in metrology). In almost all cases, a test result is a "method depending" figure.
   For example, the tensile strength of a material does not make sense if not related to the method with which it is determined.
   Consequently, the test result is conventional and no absolute or "true" value of it exists. And even when some absolute value should exist, it is practically impossible to define it with an infinite accuracy.
   For example, to determine the mass percentage of carbon in a piece of steel, it would theoretically be possible to count all atoms of the test specimen and distinguish the number of carbon atoms among the others, the number of atoms that are part of the piece of steel is likely to change constantly, for example under the effect of corrosion. Do the corroded parts belong to the steel piece or not? the answer is conventional, and this convention influences the test result;
2. Consequently, the true value should always be defined as an interval rather than a single figure. Anyways, even if a true value existed, it could never be determined exactly, what increases the width of the interval that defines it;
3. However, handling the Equation (2) with *m* expressed as an interval would not be easy. Its supposed central value is then used and called "reference value" rather than "true value" to remind constantly its conventional nature.

Comments concerning *B* and *e*:

1. To be implemented, Equation (2) requests the definition of the *B* value. This definition is provided by ISO 3534 [4] and reproduced in ISO 5725 [6] and VIM [5]. It follows that bias is related to the precision conditions in which it is determined, i.e. on the test method, on the equipment, on the operator, on environmental testing conditions. Consequently if, for example, we are considering the bias of a laboratory, we need to consider an average bias that encompasses all the methods that are implemented in the lab, all equipment used in the lab, all operators working in the lab, all environmental conditions that occur in the lab along the year. In the same way, if we are considering the bias of an operator, we need to consider an average bias that encompasses all the methods that he implements in the lab, all equipment used that he uses along the year, all environmental conditions that occur along the year when he is present;
2. To be implemented, Equation (2) requests the definition of the *e* value. Obviously, the $y - m$ term of Equation (2) is constant for a given *y* value. Consequently, this *e* value is complementary to *B* and evolves oppositely to it when the precision conditions in which it is determined are changing. Then, *e* is also related to its precision conditions.
   For example, when the effect of equipment is taken as an average in the determination of B, the random counterpart shall be included in e, but when only one test equipment is considered in B, no random effect occurs with respect to test equipment (same equipment is always used) and no contribution of it counterpart shall be included in e.

Consequences for the determination of testing uncertainties:

1. Determining properly a value of uncertainty presupposes to decide before beginning the determination what this uncertainty shall cover or, expressed with words of ISO 5725-1, which precision conditions it shall address. It can be the uncertainty on test results that are produced by a lab, or by using a defined method, or by one specific operator using one specific test equipment. In the frame of ISO/IEC 17025, the requirements are probably related to the first proposal (i.e., uncertainty on test results that are produced by the accredited lab). In most cases, determining a figure of uncertainty applicable to all test results produced by the lab is enough, but we could imagine situations where "individual" figures attached to the related "individual" test results would be useful when a very narrow uncertainty is needed by the users of the test results;
2. Depending on the decision needed about which precision conditions the determination of the uncertainty shall address, some sources of uncertainty need to be classified as sources of bias and others need to be classified as sources of random errors. Consequently, they will be expressed as mean values or as standard deviations respectively. In practice, when only one possibility occurs for a source of deviation (for example, when only one operator is qualified to perform the test method), the related effect on test results is obviously a bias. On the contrary, when many possibilities occur (for example, when a lab has got many testing machines manufactured and calibrated from different sources), the related effect on test results should be regarded as a random error. In real life of laboratories, the situations are mostly somewhere in between. Typically, including when they are many, all the testing machines of a lab are usually calibrated by a same calibration organisation, inducing a risk of bias linked to the standards of calibration used by this organisation. Consequently, in most cases, both a *B* value and an *e* value need to be considered for each source of uncertainty.

3.4 Basics for uncertainty that are specific to lab testing

Several major issues seem to us to be practically very different for the determination of uncertainties in metrology and in testing:

- How to deal with the different levels of measurands;
- The contributions of qualitative factors;
- The effect of materials;
- The inner homogeneity of the samples;
- The preparation of test specimens;
- The handling of the bias;
- The format of test results.

Because of these differences, some provisions included in GUM [2] are not adapted to the case of uncertainties in lab testing and lead to significantly underestimation of them.

How to deal with the different level of measurands

Theoretically, there is no difference between metrology and testing in this respect: in both cases, it is needed to define the range for which the uncertainty is determined and determine how uncertainty varies within this range. As a result, the uncertainty is expressed as a constant or as a percentage of the value (measured value or test result) or a combination of the two typically in the form of $U = a.V + b$.

In practice, this is usually easily performed in metrology (a same operation of calibration usually covers a wide range of measured values, typically in a ratio of 1 to 100), but requests much work, time and money for testing, especially when the method includes many qualitative sources of uncertainty.

Sometimes, the range of expected test results is narrow enough for a single experiment to cover properly the totality of the test results produced by the laboratory.

Example: When the laboratory is performing release tests of a limited range of products manufactured in a factory. In that case, it happens that the range of test results is limited enough to make possible the determination of uncertainties with a single figure.

When GUM [2] method B can be validly used, it helps to easily cover the whole range of levels of measurands encountered in practice.

Example: The linear mass LM *of a concrete reinforcing steel is determined on a piece of it which mass* M *and length* L *are measured.* LM *is then given by the equation* $LM = M/L$*. In that case, with respect to statements of this document, GUM method B can be validly used and enables to easily cover* LM *in the range from 0,1 to 10 kg/m that needs to be considered for usual concrete reinforcing steels.*

For some test methods, RM or CRM are available for a large range of test results. In those cases, a full or a partial GUM [2] method A experiment may be used to determine the impact of the range of the measurands on uncertainties.

Example: In chemistry, most of routine test methods are using CRM, which are available in many combinations of chemical composition of materials. Using several CRM enable to cover the whole range of test results produced by the laboratory.

In most of the other cases, many experiments may be needed to deal with this issue. This leads to amounts of work, time and money that are economically impossible to support. It is then recommended to explore this issue from technical knowledge concerning the test method and/or implement a partial GUM [2] method A experiment (see § 8.3).

Contributions of qualitative factors:

Theoretically, qualitative factors that affect the results have as same large importance in metrology than in lab testing. In metrology, that are issues like convection of air disturbing a mass standard or variations of the acceleration of gravity. In lab testing, it can be for example variations of pH of a corroding solution due to the corrosion of the test item. Theoretically, both are quite difficult to deal with. But in metrology, this generally affects the 6^{th} of the 9^{th} significant digit while in lab testing it generally affects the 1^{st} to the 3^{rd} significant digit. Consequently, in practice, for metrology, the issue is important only for reference calibration institutes while, in most cases, it is important in all laboratories for lab testing. And dealing with them request important resources that can be implemented in reference calibration institutes but not in average laboratories. During accreditation audits of laboratories, we have often seen laboratories that had well identified major qualitative sources of uncertainties but had considered them minor just because it would have been a huge work to deal with them.

Effect of material:

One very common “qualitative effect” is the effect of material. For many test methods, the nature of the material subject to test influences the difficulty of performing it, and consequently, the value of the uncertainty.

Example: Some metals are more subject than others to hardening during cold deformation. This cold hardening effect appears to be one to the most important sources of uncertainty during the test and consequently, values of U *depend on the type of material that is tested in addition to the other usual sources.*

Several possibilities can be implemented to explore this issue:

- Use technical knowledge concerning the test method;
- Use results from several ILC performed on several level of measurand;
- Include the effect of material in an overall GUM [2] method A experiment;
- Implement a partial GUM [2] method A experiment (see § 8.3.4).

In accordance with the amount of resources that the laboratory is ready to devote to the determination of uncertainties, several options are possible:

- Choose only one level of difficulty linked to the material, that is the average one. This option provides *U* values that are valid in most cases, but are underestimated in some cases;
- Choose only one level of difficulty linked to the material, that is the most difficult one. This option provides *U* values that are valid in all cases, but are overestimated in most cases;
- Choose to determine uncertainties for 2 types of difficulties: the average one and the most difficult one. The laboratory is then able to declare a U value that is more adapted to the tested material;
- List a series of typical materials that are tested by the laboratory and determine *U* for each of them. This option obviously provides the best estimates of *U* but is also the most requesting in terms of resources to allocate.

Inner homogeneity of the samples:

Another very common “qualitative effect” is the effect of material homogeneity. In metrology, the activities of calibration are always performed directly on calibration standards without any need of sampling or preparation. On the contrary, lab testing requests operations of sampling and of preparation of test specimens. As metrology does not need any sampling or preparation activities, the impact of them is not addressed at all in GUM [2] and consequently, many labs just ignore them or do not get a proper guidance to deal with this issue. Typically, a lab tends to consider that homogeneity of the material is not of its responsibility and take it as an excuse to avoid dealing with that difficult question. But, as a matter of fact, the lab receives a sample and select a part of it to prepare a test specimen: it is completely of its own responsibility to assure that this test specimen is faithfully representing the sample that it received. Consequently, the related effect is part of the uncertainty that it is generating.

Both calibration standards and test specimens do have an inner homogeneity SD which is not 0. But in metrology, this inner homogeneity SD is included in the uncertainty of the reference standard while it is not in the test specimen. In many cases, test methods request to perform a measurement somewhere on the test specimen and the test result depends on where on the test specimen you are performing the measurements.

> *For example, the elongation at maximum force (A_{gt}) as specified in ISO 6892-1 and ISO 15630-1 (tensile tests on concrete reinforcing steels) is measured on a given section of the test specimen. It was found by practitioners that the test result depends on the distance between breakage to basis of measurement, the length of the measurement basis and the overall length of test specimen, all of this related to metallurgical phenomena occurring inside the test specimen.*

On the other hand, some test methods may be used to investigate the global characteristic of the sample or variations of it within the samples.

> *For example, hardness tests may be used to estimate the mechanical properties of a piece of metal as well as hardness profiles of a hardened covers of mechanical pieces intended to resist to mechanical wear and tear.*

In those cases:

- When the test result is intended to represent the global characteristic of the sample, the effect related to the inhomogeneity of the sample needs to be included in the global uncertainty;
- When the test result is intended to provide information about the variations of characteristics within the sample, the effect related to the inhomogeneity of the sample shall not be included in the global uncertainty.

In the same way than before, GUM [2] does not provide any guidance about that, what may lead to underestimation of uncertainties.

Preparation of test specimens:

Most test methods request a preparation of samples that may have a major effect on test results and, for this reason, are extensively described in reference documents that describe the test method. In most cases these operations of preparation are performed by the lab or by a subcontractor under its responsibility and consequently, shall be included in the uncertainty that it is generating.

> *For example, in chemistry, the preparation of test specimens needs to avoid any kind of pollution. The effect of the unavoidable pollution within the laboratory needs included in the uncertainty.*

Handling of bias:

Metrological operations always consist in a comparison between the value of the checked item and a reference. Consequently:

1. In most cases, the bias can be evaluated (typically in the form of calibration curves) and its effect is neutralised by correcting the calibration result from this estimated bias;
2. Even when the bias component is unknown, it is always included in this comparison because it is enclosed in the uncertainty of the reference and in the results of comparisons.

On the contrary, in most cases, no references are available for testing operations so that the systematic error, i.e. bias, is not at all included in any experiment to determine the uncertainty (typically method A of GUM [2]). It follows that:

1. In most cases the bias is ignored even if it is generally the most important contribution to uncertainty;
2. Even when it is known (when some reference like a CRM is available), it is not usually possible (except in chemistry) to draw something equivalent to calibration curves, and the bias is treated as a random contribution instead of systematic one.

In both cases, the resulting effect is to underestimate the uncertainty (see § 10.9 concerning the consequence of converting a systematic deviation into a random one).

Format of test results:

Metrological results are always in a numerical form while test results may be in other forms like:

- Classification into categories (for example sweetness and aromas of wine), and whether these categories can be ordered or not (sweetness can be ordered from not sweet to very sweet, while aromas cannot);
- Binary results (for example pass/fail or presence/absence of a polluting agent) that can be regarded as a classification into 2 categories;
- Category percentages (for example classification of graphite in cast iron against ISO 945, where test results are provided in % of categories defined by standard pictures displayed in the standard);
- Curves (typically infrared spectrometry in which the IR curve is compared to a reference to identify which product the test item belongs to).

As metrological uncertainties are always related to numerical values, GUM [2] does not provide any guidance about these types of test results and testing labs usually provide very poor information about their uncertainties in those cases. Some proposals to deal with this question are then provided in § 10.1 of this document.

3.5 Classification of methods to determine uncertainties

Following the definition 2.1 of VIM [4], measurement can be regarded as a process, that have inputs and outputs. The outputs are obviously the measurement results (test results in the case of lab testing) and the inputs are all the features, conditions and parameters that influence the measurement results. Uncertainty is related to instability of the measurement process, itself linked to instability in the input parameters.

It follows that we can distinguish 2 different ways to determine uncertainties:

1. Assessment of outputs (i.e. measurement results).
   For example, processing the results of a control chart using a reference material as testing item;
2. Assessment of the impact of the variations of input parameters on the outputs of the process (i.e. measurement results).
   For example, Method B of GUM [2].

Obviously:

- Most of the usual methods are neither fully of type 1 nor fully of type 2;
- Both types have advantages and disadvantages.

Pure type 1 methods (assessment of outputs) obviously produce better estimations of uncertainties than those of type 2, because:

- There is no risk to forget an unknown source of uncertainty;
- There is no need to investigate how any variations of input parameters impact the measurement results, which is quite trickier than it looks at first observation;
- It is clear whether the bias has been introduced or not into the determination of the uncertainty.

Pure type 2 methods can be implemented by any testing lab or calibration institute alone, with no need of any collaboration with peers. For this reason and because they are well documented, they are the most widely used. However, they:

- Are usually quite more complicated to implement;
- Often request significantly higher amounts of work;
- The introduction of the bias into the determination is more complicated (taken into account in the calibration reports but not in experiments that consist in repetitions within the lab).

Consequently, they are more likely to lead to underestimated uncertainties.

In both cases, all sources of output uncertainty that are under the responsibility of the lab need to be considered in the determination of uncertainties.

3.6 Conclusions related to basics of determination of uncertainties within laboratories

The assessment of uncertainties performed during the ILC organised by CompaLab showed that most of participating laboratories significantly underestimate their uncertainties when the test method is not purely “metrological”.

The use of the GUM [2] method B is overwhelmingly dominant among all methods that can be used to determine the uncertainties, because it is well documented (in GUM [2]) and because it can be implemented without any collaboration with other laboratories.

But using GUM [2] method B to determine uncertainties leads to significantly underestimated uncertainties because it leads to mostly ignore several of the main sources of uncertainty because their impact is quite difficult to quantify. In particular, the following sources are usually badly dealt with:

- Bias because no external reference can be used;
- Preparation of samples;
- Effect of material;
- Inner heterogeneity of test specimens;
- Qualitative contributions.

It will be seen further on (see § 10.7 and § 10.8) that the main sources of uncertainty are those which make the estimation of it and since, it is of prior importance either:

- To identify and determine them;
- Or to use a method that include their effect even if they are not known.

The easy way to implement this 2nd option is to use a method that focuses on output parameters as described in § 3.5 each time that it is possible to, that is to say avoid GUM [2] method B.

Moreover, it will be seen further on (see § 4.6) that some alternate methods request less work (i.e. time and money) than GUM [2] method B.

4 How to choose properly a method of determination of U

4.1 Recommendations to choose a method of determination of U

The choice of a method of determination of uncertainties should follow the following steps:

1. Determine the level of resources that should be invested in the determination of uncertainties, with respect to the motivations for doing it;
2. Make a list of information available for the laboratory that can be re-used for determining the uncertainties;
3. Estimate, in function of the test method, the difficulty to determine properly uncertainties for the test method;
4. Make a choice of method for the main determination among the possibilities described in § 6 to § 8;
5. As function of results of issues 1 to 4 here upper, decide whether complementary studies are necessary or not;
6. Determine the uncertainties in accordance with the process decided (see § 10);
7. Check whether the determined uncertainties are consistent with the related otherwise available information (see § 11).

All these issues are discussed here after, in § 4.2 to § 4.7.

4.2 Intended use of determined uncertainties

The possible motivations for a laboratory to determine the uncertainties are several:

- Fulfil the requirements of accreditation;

- Satisfy a request of the customer of the test result (for example when the test result is intended to be used for determining the conformity of the product);
- Help analysing the sources of uncertainties in order to improve the quality of the test results provided to customers.

In the first case and the second case, a good strategy for the laboratory should be to get an acceptable result for the less possible amount of work, time and/or money. In the third case, the lab is keen to invest more in the process, what may influence the choice of the methods it should choose to determine them.

The laboratory might also follow different strategies according to the situations, i.e., for each type of testing, whether a large number of tests are performed during the year or not, whether the customers are likely to have interest in the results uncertainties or not, whether analysing the sources of uncertainties is easy or not.

When determining the uncertainties is a critical issue for the laboratory, it may use several methods and confront the results under the light of statements of § 6 to 8.

According to the answers to these questions, the laboratory should decide its level of investment in the determination of uncertainties: minimum, medium or intensive.

4.3 Definition of the field of application of the uncertainties to determine

Before beginning any determination of uncertainties, the laboratory should define the field of application the uncertainties to determine. In particular, it should be investigated to which extent the uncertainties should or not include the effects (see § 3.4) of:

- The testing methods (i.e. if the laboratory uses several testing methods);
- The inner homogeneity of the material;
- The preparation of samples.

Information should be made available to the users of the uncertainty figures about what they cover and what they do not cover.

4.4 Information available for the laboratory that can be re-used for the determination of uncertainties

Before beginning any work, the laboratory should enquire whether:

- There is available information from any ILC to which the laboratory has participated or from general documentation (typically standards);
- Some materials which central values come from an external source exist or can be produced;
- Results of programs of surveillance of quality are available in the laboratory;
- If not, does the laboratory consider taking part to an ILC and/or organising a program of surveillance of quality in the near future.

4.5 Level of difficulty to determine properly uncertainties as function of the test method

It can be seen at § 3.2 that the level of difficulty to determine properly uncertainties depend on the test method, with 3 main categories:

1. Metrological or quasi-metrological methods;
2. Methods for which both quantitative and qualitative sources of uncertainty occur;

3. Methods for which qualitative sources of uncertainty are of main importance.

In case 1, all methods including GUM [2] method B can produce good results for the determination of uncertainties.

In case 2, the method GUM [2] method B is not recommended and, if it is applied:

- Care must be taken not to miss the contribution of the bias and of the main sources of uncertainties, which are very often the qualitative ones;
- Extensive verification of the validity of determinations should be performed in accordance with § 11.

In case 3, using the method GUM [2] method B is a waste of time and money.

4.6 Choose the method to determine the main contributions to uncertainties

At this step, the laboratory is ready to choose which method is the most adapted for it. In most cases, some supplementary partial experiment or study is necessary to complete the determination among those described in § 6 to 8. Criteria for choosing are summarised in Table 2.

Table 2. Criteria for choosing an adequate method of determination of uncertainties.

Method of determination of *U*	Importance of bias	Difficulty of the test method (see § 4.5)	Efficiency	Need of resources	Comments
ILC specially designed to determine uncertainties	No	All	++	++	Not needed because of low importance of bias
	Yes	1			Worth only if the issue of uncertainties is critical
		2 and 3			Request much work, time and money but very efficient
Computation from results of participation to an ILC	Both	All	+	-	May request partial GUM method A experiment to complete the sources of uncertainty
Use of $u = s_R$	Both	All	+	--	Adapted to average labs
Full GUM method A experiment within a network of laboratories	No	All	++	+ to ++ according to the number of participants	Not needed because of low importance of bias
	Yes	All			Very efficient when coupled to the production of internal reference materials
Full GUM method A experiment using a RM or a CRM	No	All	++	- to ++	Not needed because of low importance of bias
	Yes	All			Very efficient and request: - Low additional resource when data already existing from a program of quality surveillance in the lab; - Medium resources when the test method is not destructive; - High resources (cost of RM) when the test method is destructive.
Full GUM method A experiment using internal items	No	All	++	++	-
	Yes	All	-		Requests partial at least GUM method A experiment to determine the bias

Method of determination of *U*	Importance of bias	Difficulty of the test method (see § 4.5)	Efficiency	Need of resources	Comments
Full GUM method B	No	All	+	+	May request partial GUM method A experiment to complete the sources of uncertainty
	Yes	1	+		
		2	-		The presence of qualitative sources of uncertainty makes this option not adapted at all
		3	--		

4.7 Need for additional partial GUM method A experiment

Some of the methods listed here upper make sure (if they are properly implemented) to encompass all possible sources of uncertainty (ILC specially designed to determine uncertainties, Full GUM method A experiment within a network, Full GUM method A experiment using a RM or a CRM).

In other cases, it shall be checked whether all sources of uncertainty that may occur were taken into account in the determination (typically, bias or random effect of material, method, equipment, personal, environmental conditions). Typical cases where a partial GUM [2] method A is needed and how it can be organised are described in § 8.3.

4.8 Conclusions concerning the choice of the method to determine uncertainties

It can be seen that in many cases, the GUM [2] method B is not the best method to determine uncertainties. Other methods exist, that request less efforts for better results. A series of typical situations is described here after with a better proposal for each of them:

- When information about usual s_R is available and the laboratory does not want to invest much time and money in the process, it can adopt s_R as a valid estimation of its uncertainty.
- When the laboratory has already participated to an ILC, it can use its results in this ILC to determine its uncertainties in accordance with § 6.3. This does not request much resource to invest by the laboratory.
- When information about usual s_r/s_R ratio is available, the laboratory can know whether efforts should be made rather on the determination of bias or rather on random error, or if both need to be investigated (see § 6.3).
- When the bias is of importance (almost always) and the laboratory is part of a network (for example subsidiaries of a large organisation), it should consider sharing internal reference materials by tested in all laboratories (the central value of all laboratories can then be regarded as an external source).
- When the bias is of importance (almost always) and the laboratory is not part of a network, it should enquire about existence of RM or CRM, that are then the only way to determine bias.
- When the laboratory runs a program of surveillance of quality using items which central value is known from an external source, it should use them to determine its uncertainties in accordance with § 7.4. This does not request much resource to invest by the laboratory.
- When the laboratory is considering taking part to an ILC and/or to organising a program of surveillance of quality in the near future, it should consider organising it so that he can use the corresponding results to determine its uncertainties (see § 7.4).

- If the uncertainties is a critical issue for the lab, it should consider the organisation of an ILC dedicated to the determination of them in accordance with § 6.2 and/or organise a full GUM [2] method A experiment using items which central value is known from an external source.

§ 5 to § 9 describe the methods that can be used to determine uncertainties, how they can be implemented and their respective advantages.

5 Introduction to the survey of possible methods to determine uncertainties

Here after, a survey of methods that are likely to produce acceptable determinations of uncertainties is provided with related comments. This list goes beyond the 2 GUM [2] methods that everybody sees every time everywhere, but it cannot be regarded as comprehensive. Putting once imagination on work can produce some original ideas, that can be better than the usual ones, particularly those that take advantage of some specific situations of the laboratory.

> *For example, if the laboratory is part of a research centre that has a knowledge of the product not available anywhere else, this knowledge should be used to determine uncertainties.*

In any case, all methods of determining uncertainties are kinds of experiments, which may consider all or only some sources of uncertainty according to what is technically possible and economically relevant for each situation, with regard to statements of § 4.2. Several possibilities are considered, as follows:

- Use of results of ILC specifically designed for determining uncertainties, which can be seen as a GUM [2] method A experience involving several or many laboratories;
- Use of available ILC, which requests to check whether all sources of uncertainty were taken into account;
- Organisation of a GUM [2] method A experience on reference materials (i.e. which reference value is known);
- Use of results of programs of inner quality surveillance, that can often be regarded as a GUM [2] method A experience on reference materials;
- Organisation of a full GUM [2] method A experience on materials which reference value is unknown;
- Study using the GUM [2] method B.

Whatever the method, the determination needs to:

- Take into account all sources of uncertainty (method, homogeneity, material, levels of measurand, equipment, personal and environmental conditions);
- Avoid double count of them;
- Avoid any correlation between them, or at least control it, i.e. determine its impact on the final result of the determination. It should be noted that, with respect with statements of § 10.8, the impact of a possible correlation is important only if both contributions of the correlation are among the most important.

This is particularly important and complicated to assure for studies against GUM [2] method B, but this is also important for other methods of determination.

When several series of results are available (*for example for different methods and/or for different types of test equipment and/or different operators*), the lab has got 2 possibilities:

1. Compute as many values of *U* as the number of available sets of data. A more detailed information is then available for the lab, but it then needs to manage a greater number of data when, for example, a customer requests the *U* value attached to its test results;

2. Consolidate the *U* values obtained from each set of test results into a general *U* value, valid for all test results that are produced by the lab. In that case, a weighted quadratic mean value of *U* should be used, as detailed in § 10.8.

6 Use of results from ILC

6.1 Introduction

ILC can be classified into 3 main types:

1. ILC which aim is to determine repeatability and reproducibility of a test method. The texts of reference for those are ISO 5725-2 [8] and ASTM E691 [9];
2. ILC which aim is to assess the proficiency of participants. The text of reference for this is ISO 13528 [3];
3. ILC which aim is to determine the reference value of a certified reference material (CRM). The text of reference for this is ISO 33405 (ex. ISO Guide 35) [10].

Even if the process of them is quite similar (distribution of samples as similar as possible, performance of tests by participants in given conditions, statistical processing of test results), the details of performance may be completely different, with respect with the different goals that are followed. In particular:

- The efforts of the organiser are not focused on the same issues. In case 1, attention is focused on the quality determination of repeatability and reproducibility SD. In case 2, attention is focused on the z-scores. In case 3, attention is focused on the accuracy of the central reference value. This may lead to differences in the organisation and in results of the ILC;

  For example, ILC dedicated to the determination of characteristics of CRM do not produce s_r *and* s_R *that represent the accuracy of the test method implemented by an average laboratory, because the participants to such ILC are usually very skilled laboratories, with uncertainties lower than the averages.*

- The types of participants are different. In case 1, participants are peers of the ILC organiser, generally from the academic world. In case 2, participants are customers of the PT provider. In case 3, participants are subcontractors of the ILC organiser. The level of skill of participants and the control of the organiser over them is different in the 3 cases;
- The types of items are different. In case 1, CRM may be used to assure a better homogeneity of them. In case 2, items are supposed to represent the day-to-day life of participating labs, i.e. industrial items. In case 3, items are intended to become CRM but are not assessed yet;
- Levels of quality assurance of the organiser are likely to be different. In case 1, no specific surveillance is required while cases 2 and 3 are subject to accreditation, i.e., (voluntary) external surveillance.

This list is not comprehensive, but it clearly shows that level of confidence in the information coming from ILC can vary a lot, according to the sources and the parameters that are used to determine uncertainties.

Even if we never encountered such an exercise, it would make sense to organise an ILC specific to the determination of uncertainties, in which all decisions for the organisation of it (in particular the options for precision conditions as defined in ISO 5725-1 [6]) would be focused on that goal.

In any cases, ISO 21748 [11] provides guidance on how to deal with ILC results in order to determine uncertainty values. This standard is based on 2 principles as follows:

1. The standard deviation of reproducibility is a valid basis for the determination of the uncertainty;

2. The sources of scatter not included in the ILC program shall be identified and added if it cannot be demonstrated that they are negligeable.

It identifies some sources of scatter that, typically, may be not included in the ILC scheme:

- A bias linked to the test method;
- The effect of the material (is the material different from those usually encountered in the laboratory);
- The level of homogeneity (whether or not the internal homogeneity of the sample should be taken into account, see § 3.4);
- The level of the measurand, if it is significantly different from those usually encountered in the laboratory;
- The preparation of the test specimens when they are prepared by the organiser of ILC.

This happens when the organiser of the ILC requests the participants to use a given method and/or when it prepares itself the test specimens. These conditions normally depend on the aim of the ILC are likelier to happen in ILC of type 1 and 3 than in type 2.

Here after, 3 different issues will be discussed:

- What could be an ILC specifically designed to determine individual uncertainties;
- How results of an ILC can be used by a participant to determine its uncertainties;
- How global results of an ILC can be used as reference values for uncertainties.

6.2 ILC specifically designed to determine individual uncertainties

If an ILC specifically designed to determine individual uncertainties were organised, it should include all sources of uncertainty that impact test results, i.e.:

- Possible bias related to the test method;
- Possible bias and random error related to the material and its homogeneity, as function of the intended use of the test results (see § 3.4);
- Bias and random error related to the levels of measurand;
- Possible bias and random error related to the preparation of test specimens;
- Bias and random error related to the performance of the tests.

For achieving that, the ILC scheme should include:

- Enough test results from all methods that can be used to determine the test results;
- Preparation of test specimens by participants, so that the possible related bias and random error is included in the interlaboratory SD and in the global bias of the lab;
- Performance of tests include precision conditions (see § 3.3) that cover all the sources of uncertainty encountered in practice within each lab, i.e. its whole testing equipment, all its staff and all the environmental conditions that can be encountered along the year.

ISO 5725-3 [12] (that defines alternate schemes for precision tests) can be used to determine adequate ILC schemes to achieve these goals.

Using these ILC conditions, the statistical processing of data can determine:

- The bias related to each testing method. When one test method is regarded as reference, the reference value shall be computed from the related results. When not, the reference value should be taken as the average of all test results;

- The true value of the bias related to each lab;
- The true value of the random error of each lab.

From these results, each lab can determine the value of the enlarged uncertainty *U* related to the test results it produces using Equation (3):

$$U = |X_{Lab} - X_{ILC}| + k.\sqrt{u_{XILC}^2 + s_{Lab}^2} \quad (3)$$

where U *is the enlarged uncertainty of the test result,*
X_{Lab} *is the mean value of the test results of the lab,*
X_{ILC} *is the reference value computed from the totality of the ILC test results,*
k *is the enlarging coefficient,*
u_{XILC} *is the uncertainty on the of the* X_{ILC} *value,*
s *is the standard deviation of the test results of the lab.*

A term B_{met} representing the bias of the method may be added if it is not included in X_{Lab}.

Conclusion concerning organisation of an ILC specifically designed for determining uncertainties:

This possibility is as probably the most powerful method to determine uncertainties because all sources of them are by design included in the experiment. However, it requests a whole organisation of it, and consequently, is costly in time and money.

This possibility could be adapted when determining accurate values for uncertainties is critical.

6.3 Use of results of ILC that was not specifically designed for determining uncertainties

6.3.1 Introduction

ISO 21748 [11] provides extensive information on how to use ILC results for determining uncertainties. It says that the reproducibility SD is a good basis for the estimation of the standard uncertainty.

When the impact of the test method and/or of the preparation of test specimens are not included in the scatter estimated in the ILC, it should be added. In both cases, this can be achieved by a partial GUM [2] method A experiment as explained in § 8.3.

6.3.2 Computation of individual participants uncertainties

As the same model (i.e. Equation (2) from [6]) is used whatever the type of ILC, Equation (3) can also be used to compute the uncertainty on results of an individual participant. We have seen at § 6.2 an ideal configuration of ILC for which the totality of sources of uncertainty are included in the calculations. In most cases, usual ILC have slight discrepancies to this situation which typically are related to:

- Representativity of mean values and standard deviations;

  For example, an ILC that was performed to determine the characteristics of CRM is likely to provide underestimated s_r *and* s_R *values, because the laboratories that took part to it are likely to be more skilled than average laboratories.*

- Test method, when several test methods are possible, and it was requested to use a method which is not the method of reference;
- Test specimen preparation, when such a preparation is necessary and ILC was conducted on test specimens prepared by the organiser;

- Precision conditions (as defined in [6]). In most cases, participants are requested to perform the tests in repeatability conditions. Consequently, impacts of test equipment, operators and environmental conditions are not included in the estimation of random error.

When such situations occur, it is needed to add one or several contributions as follows:

- The bias related to the test method;
- The bias and random error related to preparation;
- The bias and random error related to test equipment, operators and environmental conditions.

When the impact of the test method and/or of the preparation of test specimens are not included in the scatter estimated in the ILC, it should be added. In both cases, this can be achieved by a partial GUM [2] method A experience as proposed in § 8.3.

When such impacts need to be added, Equation (4) for computing *U* applies, as follows:

$$U = \left|X_{Lab} - X_{ILC} + B_{Met} + B_{Prep} + B_{EPC}\right| + k.\sqrt{u_{XILC}^2 + s_{Lab}^2 + s_{Prep}^2 + s_{EPC}^2} \quad (4)$$

where U *is the enlarged uncertainty of the test result,*
X_{Lab} *is the mean value of the test results of the lab,*
X_{ILC} *is the reference value computed from the totality of the ILC test results,*
B_{Met} *is the bias related to the method,*
B_{Prep} *is the bias related to the preparation,*
B_{EPC} *is the bias related to test equipment, personal and environmental conditions,*
k *is the enlarging coefficient,*
u_{XILC} *is the uncertainty on the of the* X_{ILC} *value,*
s_{Lab} *is the standard deviation of the test results of the lab,*
s_{Prep} *is the standard deviation related to the preparation,*
s_{EPC} *is the standard deviation related to precision conditions in which the ILC was conducted.*

When s_{EPC} includes a SD of repeatability, s_{Lab} should then be omitted to avoid double counting of this impact.

These calculations can be performed either by the ILC provider or by the participant from the information provided in the ILC report.

On its own, CompaLab computes uncertainties by this method for its customers. ILC organised by CompaLab are of type 2 of § 6.1. Consequently, the policy for designing them is to:

1. Have all operations that are part of participants proficiency performed by participants;
2. Put the participants in testing conditions as close as possible to their usual ones.

Then, in ILC organised by CompaLab:

- All methods enabling the determine the expected result are allowed. Participants should use those methods that they are used to;
- Preparation of test specimens shall be carried out under their responsibility.

Consequently, the 2 main issues addressed at § 6.1 are fulfilled.

Moreover, when several methods are possible and enough participants provide test results for a given test method, a separate statistical processing is performed for this method, enabling to check whether:

- This method causes a significant bias;

- Any possible alert is due to a poor performance of the method or to the method itself or a combination of the two.

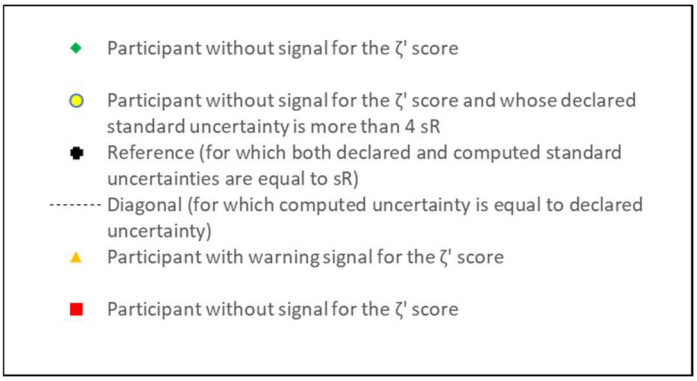
  In practice of CompaLab ILC, comparisons of methods typically apply to chemical tests, determination of thickness of coatings, methods of controlling the loading rate in mechanical tests and methods of determination of elongation at maximum force A_{gt} *on concrete reinforcing steels. These comparisons are performed using the classical formula* $|X_{Met} - X_{Overall}|/\sqrt{u^2_{XMet} + u^2_{XOverall}}$, *which should not overcome 2. In almost all cases, no significant differences can be out in evidence except for* A_{gt} *for which results from ISO 6892-1 are lower than those from ISO 15630-1 by a significant average ratio of 0,9.*

However, participants are requested to perform the tests in repeatability conditions, to make possible the assessment of their repeatability and to avoid that the performance of the ILC does not take too much time. Because of this, a s_{EPC} SD should be added to the computation of U, but the corresponding information is not available for doing it. Consequently, this term is ignored and should be added by the participants.

Despite this problem, the uncertainties computed by this way are quite good, as shown in Figure 2.a to c.

These figures provide comparisons of u_d declared by participants, u_c computed by CompaLab and s_R (reproducibility SD) for ILC which were used in Figure 1. They are built up as follows:

- Abscissas are u_d, standard uncertainties declared by participants;
- Ordinates are u_c, computed with Equation (4) from results of ILC. It must be noted that Equation (4) provides enlarged U uncertainties and not standard u ones. To make comparisons possible, we considered that k is almost always chosen equal to 2, and consequently, we chose to compute $u_c = U_c/2$, that is to say include half of the bias contribution, even if it is obviously theoretically not correct;
- With respect to statements of ISO 21748 [11], s_R is taken as reference, i.e. is plotted as [100%;100%];
- Scales of axis are logarithmic in %, to cope with huge differences that usually occur between the different types of determination;
- Caption is as follows:



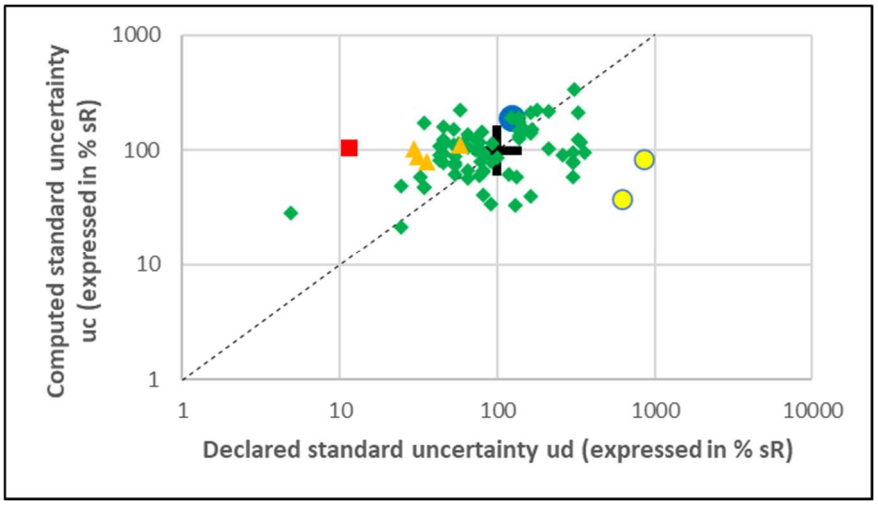


Figure 2.a: Results of 2023 of assessment of uncertainties for carbon content on low alloyed steel.

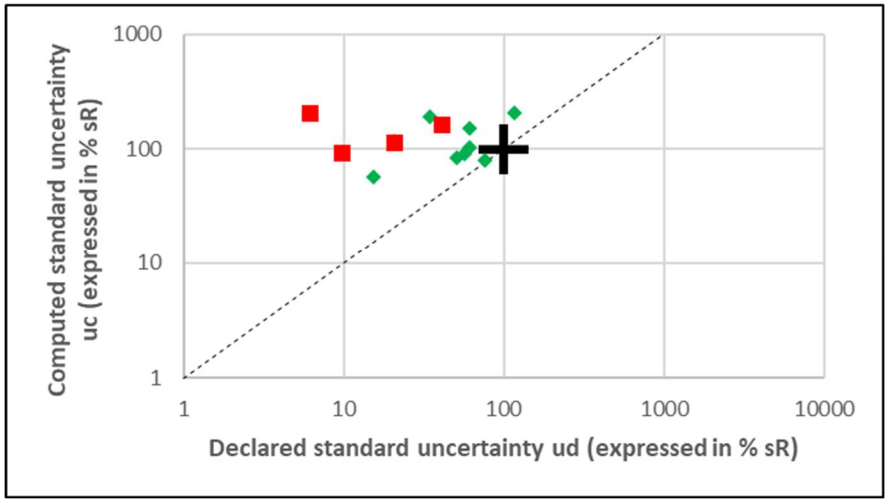


Figure 2.b: Results of 2023 of assessment of uncertainties for A_{5d} (elongation after rupture) in a tensile test on a carbon steel.

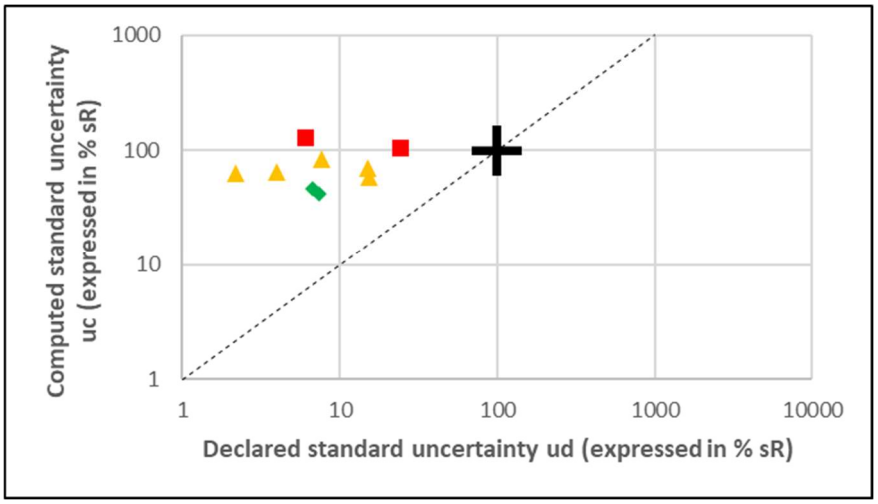


Figure 2.c: Results of 2023 of assessment of uncertainties for speed of intergranular corrosion in a Huey test on a stainless steel.

It can be seen from these Figure 2.a to c that:

- When uncertainties are globally well determined (Figure 2.a, corresponding to case 1 of § 3.2), both u_c and u_d are located in the same range than s_R. However, the scatter of u_c values is lower than the scatter on u_d values, letting us think than u_c are more accurately determined than u_d;
- When uncertainties are globally not well determined (Figure 2.b and c, corresponding to cases 2 and 3 of § 3.2), u_c is located in the same range than s_R while u_d values are all lower than u_c and almost all significantly lower than s_R. In case 3, u_d/s_R ratios are between 3% and 30%! Moreover, the scatter of u_c values is also significantly lower than the scatter on u_d values, letting us think than u_c are more accurately determined than u_d;
- Alerts from ζ-scores are consistent with u/s_R and u_d/u_c ratios.

Conclusion concerning computation of individual participants uncertainties:

Each time when ILC exist, using their results is a good solution to compute uncertainties, as most of the work is made by external sources. However, it needs to be checked whether some terms are missing and sometimes, it requests additional partial GUM [2] method A experiment to address them. Consequently, it can sometimes lead to underestimations when it is technically or economically complicated to determine missing terms of Equation (4).

6.3.3 Use of reproducibility SD as an estimate of uncertainty

As stated in ISO 21748 [11], s_R is a good basis for estimating uncertainties, as main sources of uncertainty are usually included in the experiment and as possibly lacking terms are well identified (see here upper).

Moreover, many reasonably reliable information is freely available about them, as for example:

- Standards, especially ISO standards concerning chemistry and many ASTM concerning test methods, include annexes that contain results of experiments of precision against ASTM E691 [9] or ISO 5725-1 [6], see annex;
- Scientific literature concerning the performance of many types of testing.

However, u values obtained by using this way are relevant for average laboratories, i.e.:

- Neither very skilled laboratories, where important efforts are made to reduce the uncertainties of the test results;
- Nor poorly skilled laboratories. ISO 21748 [11] actually states that it shall be verified that the bias of the lab is under control (i.e. within what is expected in an ILC).

It is always possible to check whether the bias of a lab is under control by using a partial GUM [2] method A experiment as explained in § 8.3.

Conclusion concerning the use of reproducibility SD as an estimate of uncertainty:

This very easy method (just look for free information) is producing an acceptable reference value for uncertainties in averagely skilled laboratories, i.e. 90% of them.

6.3.4 Use of repeatability and reproducibility SD to estimate the relative importance of bias and random error

In most configurations of ILC:

- The standard deviation of mean values of participants s_L is related to the average bias of participants;
- The standard deviation participants results s_r is related to repeatability;
- A relationship exists between them and the reproducibility standard deviation s_R, in accordance with the Equation (5) from ISO 5725-2 [8], as follows:

$$s_R = \sqrt{s_L^2 + s_r^2} \quad (5)$$

Consequently, the ratio s_r/s_L and the mathematically related ratio s_r/s_R provide information on the relative importance of bias and random error for a given test method. Moreover, in the same than in § 6.3.3, these values are available from many sources. We can distinguish 3 typical different situations:

- When $s_r/s_R < 0{,}2$. Then, the bias is of major importance and efforts should be focused on it;
- When $0{,}2 < s_r/s_R < 0{,}9$. Then, both bias and random error may be significant and should be considered;
- When $s_r/s_R > 0{,}9$. Then, the bias is of minor importance and efforts should be focused on random error.

This last situation occurs when the issue of inner homogeneity of the material subject to testing is of major importance for the determination of uncertainties.

In CompaLab ILC, the first type of situations usually occurs for chemical testing, the last situation usually occurs for most of the tests on concrete reinforcing steels. The second type of situations usually occurs for almost all the other test methods.

Conclusion concerning the s_r/s_R ratio:

Very low or very high s_r/s_R ratios enable the laboratory to avoid spending time and energy in determining the contributions of respectively random error and bias:

- When the random error is not significant compared to bias, a single partial GUM [2] method A experiment intended to determine bias (see § 8.3.2) is enough to determine properly the uncertainty of the laboratory;
- When the bias is not significant compared to random error, there is no need to look for external sources for reference values (i.e. RM or collaboration with other laboratories).

7 Experiment according to GUM method A for which test items are a RM

7.1 Introduction

Experiments according to GUM [2] method A consist in performing a large number of tests for which precision conditions (as defined in ISO 5725-1 [6]) are controlled. This control of precision conditions enables to determine the random error. If, moreover, it is conducted on a RM (for which the reference value is known), the bias can also be determined, and a full determination of enlarged uncertainty is possible.

This experiment is of type 1 according to § 3.5, i.e. the assessment is focused on outputs of the testing process.

In the same than before, the following issues should be considered for the precision testing:

- Method;
- Material and its homogeneity;
- Levels of measurands;
- Preparation of test specimens;
- Testing conditions (test equipment, personal, environmental conditions).

In practice, these experiments can be performed:

- By organising an overall experiment within the laboratory, that should cover all inner sources of uncertainty;
- By using available data within the laboratory. Typically, overall results of a possible program of inner quality surveillance can be used.

7.2 Reference values need to come from an external source

Obviously, the bias (which is the systematic deviation of the lab) cannot be estimated with an internal experiment of accuracy (typically GUM [2] method A or B). Consequently, it is a major importance that the reference value X_{RM} comes from an external source. However, it does not need to be a CRM (Certified reference material), even if CRMs are usually the best external sources that can be found. Beside CRMs, external sources can be:

- Items that were used in an ILC, which report provides the corresponding central reference value X_{RM}. However, this option can only be used either when the test method is not destructive or if some items to be tested remain after the end of the performance of the ILC;
- Items which central reference value X_{RM} can be known from another source, which can be internal or external;

  For example, for the determination of cement content in hardened concrete (which is performed with a picture analysis method), it is possible to produce internally samples which cement content is accurately known from its parameters of preparation;

- Items that were internally prepared for that purpose and tested by several laboratories;

  For example, a large company that has several subsidiaries in several places may prepare a set of samples that are used for quality surveillance in each place. When a large quantity of test results from all the testing places is available, the whole data can be used to determine a reference value.

If the reference value does not come from an external source, the estimation of U omits the bias, which leads to significant underestimation of it (see § 10.8).

7.3 Organisation of an overall experiment within the laboratory

The first step for such an experiment is to list all sources of uncertainty that are encountered within the laboratory and then, produce a design of experiment that encompasses all these sources. As in other types of experiments, the design of experiments must represent as faithfully as possible the sources of uncertainty that influence the test results of the laboratory, not more, not less.

For example, if a laboratory prepares itself the test specimens, uses 2 testing methods, 8 testing machines and 5 operators, the design of experiments should include:

- *2 testing methods;*
- *2 types of material (average and difficult);*
- *5 levels of measurands;*
- *3 sources of preparation of test specimens (its own facilities plus, as far as possible, 2 external sources);*
- *8 testing machines;*
- *5 operators;*
- *2 environmental conditions (i.e. performance of tests at 2 different periods of the year);*
- *2 repetitions, if repeatability SD is needed.*

That is to say a total number of 2x2x5x3x8x5x2x2 = 9600 tests.

In almost all cases, this represents a quantity of tests that is technically and/or economically impossible to support. However, it is possible to reduce drastically this number by the following means:

- In most cases, not all combinations apply in practice. The design of experiments should represent only the actual practice of the lab. In particular, when correlations occur, they should be taken into account in the design of the experiment;

For example, in most cases, when a lab is operating 2 testing methods, 8 testing machines with 5 operators, each of the operators is operating only 2 or 3 or 4 machines, not all machines and not all operators can produce results with the 2 methods.

- The experiment may be split into several partial experiments, that deal separately with each of the issue. However, possible (favourable or unfavourable) interactions between sources may then be omitted;

For example, the impact of preparation may be checked for each of the methods but with only 1 machine and 1 operator.

- A random design of experiments may be used, where the checks are distributed to cover a random selection of configurations among those actually practiced within the laboratory (see an example in annex);
- These partial experiments or random design of experiments may include some tests that are performed on RM and some others that are not performed on RM.

For addressing the 1st issue, a list of configurations actually used in the lab must be established.

Using the 2nd option requests to make sure that no interaction between factors occurs. In that case, the design of experiments should take it into account.

For example, the impact of preparation may be different according to the machines that are used.

When possible (i.e. when enough knowledge about the issue is available) it can also be decided to choose so kind of medium configuration, representing the average effect of preparation.

Using the 3rd option needs to make sure that the random scheme covers properly the actual configurations used in the lab. In particular, the total number of tests shall be too much reduced and no important source of uncertainty should be avoided in the testing scheme.

For example, if one of the 8 testing machines is used to produce 40% of test results, the random scheme should make it out of the experiment.

When a great number of configurations is actually used, it is possible to weight them with the frequency with which they are used, to produce the random design of experiments.

An example of full design of experiments is provided in annex.

Conclusions concerning the organisation of an overall GUM [2] method A experiment within the laboratory, using reference materials:

When they are designed properly, such experiments provide very good estimations of uncertainties but they request the existence of RM and are costly in time and money.

7.4 Use a statistical processing of test results performed for quality surveillance

ISO/IEC 17025 [1] (which is the reference for lab accreditation) requires labs to conduct a quality surveillance that, among other possibilities, can be control charts. On the other hand, some standards describing test methods (for example ISO 6507-1 [7]) also request to check periodically the testing machine with a RM. This produces a set of data that can be used to determine uncertainties. Figure 3 provides an example of such a data.

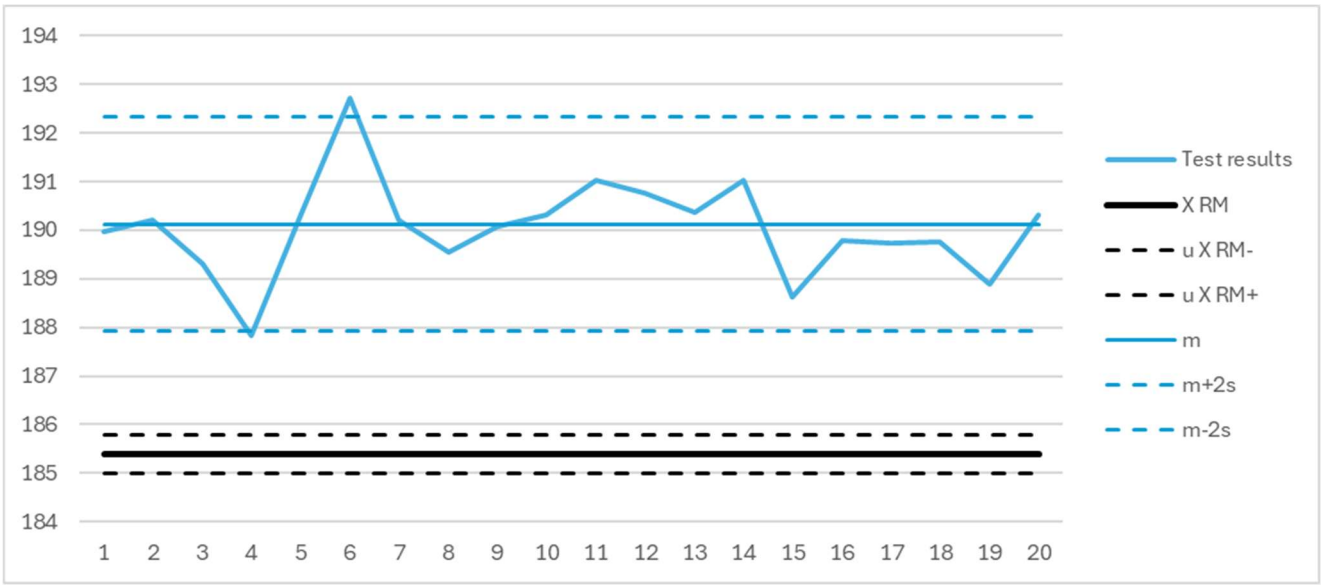


Figure 3: Example of control chart which results can be used to determine the uncertainty on test results.

In this example, it can be seen that:

- $m - X_{CRM}$ where m is the mean value of the test results on the RM and X_{CRM} is the assigned value of the RM, is a good estimator of the bias B;
- s is a good estimator of the random error e.

It shall be also noted that the chart of Figure 3 does not look like usual control charts as described in the ISO 7870 series. Typical control charts are normally centred on the reference value and limits of alert are normally computed and displayed on the chart. It is important to note that the aim of usual control charts is to control drift and not bias and scatter. Consequently:

- The central value of a usual control chart is determined internally and represents a “zero” point of the state of the process. If any drift occurs, it can be seen as a deviation to this “zero” point;
- A usual control chart cannot be built up only by using external reference materials because its limits of IC normally do not represent the acceptable deviation to the reference value (either they are definitely too severe, or the RM is very bad).

In small laboratories, the control charts of surveillance of quality of testing usually includes test results from different test machines and different operators, each of them being alternatively checked. But in large laboratories or when the quality of test results is very critical, several or even many control charts may be implemented, each of them being devoted to a specific machine or a specific operator. Obviously, in the first case, the computed B and e parameters are those of the whole lab, while in the second case, the computed B and e parameters are those of the corresponding test equipment and/or operator. Consequently, the corresponding U may be different from one to the other control chart.

Conclusions concerning the use of a statistical processing of test results performed for quality surveillance:

When such data exist within the laboratory, it should absolutely be used to determine uncertainties because the basic method is very effective and the additional time and money necessary to achieve it is very low.

7.5 Computation of *U*

The enlarged uncertainty U can then be computed with Equation (6), as follows:

$$U = |m - X_{RM}| + k.\sqrt{u_{RM}^2 + s^2} \quad (6)$$

where U *is the enlarged uncertainty of the test result,*
m *is the mean value of the series of test results,*
X_{RM} *is the assigned value of the RM (reference material)*
k *is the enlarging coefficient,*
u_{RM} *is the uncertainty on the of the* X_{RM} *value of the reference material,*
s *is the standard deviation of the series of test results.*

In the example of Figure 3, $U = |190{,}0 - 185{,}4| + 2.\sqrt{0{,}4^2 + 1{,}0^2} = 6{,}8$, which encompasses well the gap between any of the test results and the reference value.

It can be argued that *m* and *s* are estimators which can be a bit distant to the true values that are supposed to represent. To cope with this, it possible to include in the term representing *e* an enlargement by using a *t* value coming from the Student tables instead of the usual *k* enlarging coefficient. This is however contradictory with the first principle of GUM, which states that determinations of *u* should be as realistic as possible rather than determining by excess.

This was confirmed by the Monte-Carlo method (see § 10.11), which results are as follows:

- *U* = 6,819 ± 0,006;
- 95% of *U* estimations belong to the interval [5,99;7,65].

7.6 Comments concerning *B* and *e*

Following statements of § 3.3, it is needed to distinguish what is included in the *B* term and what is included in the *e* term during the quality surveillance experiment.

If the data of test results comes from a organised experiment that was conducted within a short period of time, the *e* term may be lower than when the results of a control chart is used, for which results obtained within a long period of time.

In most cases, RM are provided in the form of test specimens. Then, the results used in both situations (organised experiment and control charts) do not include the impacts of preparation of test specimens. Consequently:

- The contribution of preparation of test specimen of RM appears as bias in this type of uncertainty experiment;
- This bias is taken into account in the u_{XRM} value;
- The experiment does not include the contribution of preparation of test specimens in the day-to-day life of the laboratory, and this contribution should be estimated separately, for example by a partial experiment in accordance with GUM [2] method A, see § 8.3.6;
- RM usually have better inner homogeneity than day-to-day test specimens. The corresponding impact on uncertainty might then be underestimated.

In the same way, when several methods may be used to get the test results, it shall then be checked whether the RM is valid for only one method or not and decide in consequence whether a contribution of the method to uncertainty is needed or not.

In any case, it is needed for each situation to analyse carefully what is included in the *B* term and what is included in the *e* term.

7.7 Conclusions concerning the experiments according to GUM method A with RM as tested items

Experiment conducted in accordance with GUM [2] method A using RM as tested items is usually an efficient type of experiment because it usually does not forget any major source of uncertainty. However, it shall always be checked whether contributions (bias and/or random error) from preparation of test specimens, material and inner homogeneity of test specimens (bias and/or random error), level of measurands and possible bias of the test method need to be added or not.

When results from inner program of quality surveillance are available, they should always be used as data for such experiments because the determination of *U* becomes then easy and cheap to undertake (one mean value and one standard deviation to compute). However, in many cases, it is difficult or expensive (when the test method is destructive) to find an external source for the reference value X_{RM}.

In other cases, it is always possible to organise an overall experiment for which the organiser can control all parameters, and consequently, address all sources of uncertainties, in particular the bias. However, it can be technically or economically difficult to get RM in enough quantity and it can request much time and money to perform the tests requested to get the results. As a conclusion, organising such an experiment is relevant only when it is needed to provide accurate values of *U*.

8 Experiments according to GUM method A for which test items are not RM

8.1 Introduction

When no RM can be made available, it always possible to organise GUM [2] method A experiments, which can be:

- Overall experiments and/or use of results from programs of inner quality surveillance;
- Or partial experiments, that deal with a specific source of uncertainty, complementarily to other experiments.

8.2 Overall experiments and/or use of results from programs of inner quality surveillance

Overall experiments performed in accordance with GUM [2] method A on test items which reference value is not known from an external source do not make possible the estimation of bias, and consequently, ignore a major source of uncertainty in most of the cases.

For example, in the example of Figure 3, U *= 6,8 when the bias is taken into account and* U *= 2,2 when it is not.*

Consequently, this is a valid method only when:

- The bias of the lab can be estimated by a separate partial experiment, see § 8.3.2;
- Or when it can be proved that no significant bias applies to that type of test, see § 6.3.4.

In that case, recommendations of § 7 and 8.3.8 are applicable.

Otherwise, such an experiment is in fact a partial experiment as described in § 8.3.

8.3 Partial GUM [2] method A experiments

8.3.1 Introduction

Partial GUM [2]method A experiments can be used to determine the impact of one or several identified sources of uncertainty but not all of them together. In particular, such partial experiments may be used to determine contributions that are difficult or impossible to determine by other means, such as:

- The bias;
- The test method;
- The effect of material;
- The range of measurands;
- The preparation of test specimens;
- The precision conditions as defined in ISO 5725-1 [6].

8.3.2 Partial GUM [2]method A experiment for determining the bias

To determine the impact of the bias, a GUM [2] method A experience can be organised as follows:

- Get a series of RM. It is not needed to be a CRM. Other possibilities listed in § 7.2 can also be used. When no RM is available from external sources, it is needed to prepare them internally and make them tested by an enough number of labs, independent from each other (see § 10.7 to determine the minimum number that should be used);
- Compute the average value of each lab and the overall average. SD values may also be determined for information;
- Compute the bias related to the lab for each of the levels of performance.

8.3.3 Partial GUM [2]method A experiment for determining the impact of the test method

To determine the impact of the test method, a GUM [2] method A experience can be organised as follows:

- Prepare an enough number of samples as similar as possible;
- Perform the test according to each method on an equal number of test specimens;
- Compute the reference value from the results obtained with the reference method if any, from all the results if not;
- Compute the overall SD values;
- Compute the bias and the SD related to test method that are usually used by the laboratory.

8.3.4 Partial GUM [2]method A experiment for determining the impact of the material submitted to testing

To determine the impact of the material submitted to testing, a GUM [2] method A experience can be organised as follows. In accordance with the amount of resources that the laboratory is ready to devote to the determination of uncertainties, several options are possible:

- Decide one or several materials corresponding to different levels of difficulties for performing the tests (see § 3.4);
- Prepare an enough number of samples of each material, as similar as possible within each material;
- Perform the tests within a single laboratory on all test specimens;
- Compute the averages and SD values for each level of measurand;

- Allocate uncertainties values for each of them.

8.3.5 Partial GUM [2]method A experiment for determining the impact of the range of test results

To determine the impact of the range of test results, a GUM [2] method A experience may be organised as follows:

- Prepare an enough number of samples of at least 3 levels of measurand, as similar as possible within each level of measurand. According to the importance of the issue, 3 to 10 levels should be tested. These levels should be chosen among preferred numbers as described in ISO 3 [13] (i.e. 1 – 1,25 – 1,6 – 2 – 2,5 – 3,15 – 4 – 5 – 6,3 – 8 – 10 – etc. …);
- Perform the tests within a single laboratory on all test specimens;
- Compute the averages and SD values for each level of measurand;
- Plot these results as function of the levels of measurands and compute a relationship between the two (typically in the form of $U = a.V + b$).

8.3.6 Partial GUM [2]method A experiment for determining the impact of preparation of samples

To determine the impact of the preparation of samples, a GUM [2] method A experience can be organised as follows:

- Prepare an enough number of samples as similar as possible;
- Ask several different entities to prepare an equal number of test specimens;
- Perform the tests within a single laboratory on all test specimens;
- Compute the overall average and SD values;
- Compute the bias and the SD related to preparation for the entities that usually prepare the test specimens for the laboratory.

8.3.7 Partial GUM [2]method A experiment for determining the impact of the precision conditions

To determine the impact of the precision conditions (as defined in ISO 5725-1 [6]), a GUM [2] method A experience can be organised as follows:

- Prepare an enough number of samples as similar as possible;
- Perform the test on an equal number of test specimens, using all test equipment, operators over a period of time long enough to represent the variety of environmental conditions encountered in the lab along the year;
- If the determination of the repeatability SD is needed, each configuration shall be tested at least 2 times;
- Compute value the repeatability SD value and the overall SD according to Equation (7).

$$s_r = \sqrt{\left(\sum_i s_i^2\right) \Big/ (n-1)} \qquad (7)$$

where s_r *is the estimated repeatability SD,*
s_i *are the individual standard deviations computed on each individual test configuration of test equipment, operator and environmental conditions,*
n *is the total number of test configurations.*

The proposal here upper include all contributions to random error within the laboratory. It is of course possible to limit the experiment to some contributions only.

8.3.8 Updating U with results of partial GUM [2] method A experiments

Equation (4) can be used to complete the calculation of *U*. When the results of Equation (4) are known in a consolidated way, the addition of supplementary terms coming from a partial GUM [2] method A experiment shall not be implemented in the same for bias and for the random error.

Bias:

The part representing bias in Equation (4) is as follows: $B = |\sum B_i|$. The addition a supplementary term B_n to *B* may increase or decrease it, depending on whether B_n is of same sign than $\sum B_i$ or not. Consequently:

- When $\sum B_i$ is known, updated *B* should be computed as $B_{updated} = |\sum B_i + B_n|$;
- When only *B* is known, updated *B* should be computed as $B_{updated} = B + |B_n|$, which may be overestimated.

Random error:

The part representing the random error in Equation (4) is as follows: $e^2 = \sum e_i^2$. The addition a supplementary term e_n to *e* is just adding an e_n^2 term to the sum $\sum e_i^2$. Consequently, in all cases, $e_{updated}^2 = e^2 + e_n^2$.

8.3.9 Conclusions concerning partial GUM [2] method A experiments

These partial experiments do not enable to investigate interactions between each of the sources of uncertainty. Consequently, they should be regarded as complementary to an overall experiment when it is not technically or economically possible to include them into the main one.

They cannot provide a reliable value of *U* but can be very useful to complete the calculations when information is available from elsewhere concerning some contributions but not all.

In many cases, not enough external information is available and a full GUM [2] method A experiment is technically or economically very difficult or even impossible to organise. In those cases, they are really useful for determining properly *U*.

9 Studies according to GUM method B

9.1 Introduction

The GUM [2] method B differs significantly from all the upper because it is not mainly based on experiments but on an analysis of causes of uncertainty and combination of the effects of each of these causes, according to Equation (8), as follows:

$$u = \sqrt{\sum_i u_i^2} \qquad (8)$$

where u *is the estimated global uncertainty,*
u_i *are the individual values of effects of each cause of uncertainty.*

Comments about this method:

- It is then clearly a method of type 2 according to the classification of § 3.5, which supposes to transform the uncertainties on causes into uncertainties on effects;

- It can also be seen that the bias cannot be handled separately with this method, and it needs to be handled as a random effect among others;
- This method does not necessarily request the performance of experiments. Much information (typically all uncertainties related to metrological measurements) can come from documents or any relevant external information.

9.2 Implementation of GUM method B

The following operations are needed to implement it:

1. List accurately all sources of uncertainty that occur during the testing process;
2. Evaluate the uncertainty on each of these input parameters;
3. Evaluate the individual effect of them onto the final result;
4. Compute the combination of all effects.

Listing the sources of uncertainty:

This list needs to be:

- Comprehensive: if some sources of uncertainty are omitted, Equation (8) will lack some terms and the final *u* result will be underestimated. Many methods can be used to make it as comprehensive as possible. Typically, the analysis should consider the sources linked to method, material, homogeneity, levels of measurands, preparation of test specimens, equipment, personal, environmental conditions. To cope with these items, it is often useful to check the standard describing the test method: obviously, all parameters that are needed to be controlled are those that influence the test result. Consequently, any small deviation of them are likely to cause a small deviation in the test result. Scientific literature can also be used to make the list as comprehensive as possible. However, it can be difficult to achieve comprehensiveness because some sources may remain unknown;
- Without double count: if some sources of uncertainty are double counted, Equation (8) will contain several times the same effect and the final *u* result will be overestimated. However, the practice shows that it can be difficult to think about some hidden double counting;

  Example 1: When a global evaluation of the effect of the test equipment is achieved and when that global evaluation includes the bias, the uncertainty linked to the calibration of related sensors shall not be considered as another source of uncertainty.

  Example 2: When a calibration report of a sensor is used, it includes the contribution of the resolution of it. Consequently, no contribution of the resolution shall be added.

  Example 3: When a mark is manually applied on test specimens (to measure elongation during tensile tests) and the effect of operators is determined by a method A experiment using several test specimens, the effect of this manual operation shall not be added because it is already included in the method A experiment.

- With controlled correlation between themselves. Equation (8) is valid only if all sources of uncertainty are statistically independent, i.e. if, in practice, no relationship occurs between the terms. Checking this statistically is a huge work that nobody never does. The easy way to check it is to consider it on the technical point of view. When such correlations are likely to happen, the best way is to evaluate of the related effects together in a global analysis or GUM [2] method A partial experiment (see § 8.3).

For example, when same test equipment is always used by same operators, a correlation is likely to happen between equipment effect and operator effect. It is then advised to consider a coupled equipment-operator effect that include both effects and their correlation.

Evaluating the uncertainty linked to each of these input parameters:

The uncertainty linked to each of the identified input parameters needs to be estimated. It may come from:

- Requirements of the test method;
- Results of former experiments, including results from former participations of the lab to ILC;
- General experience or knowledge of the properties of the materials and instruments that are used;
- Specifications of the manufacturer;
- Data from calibration or other certificates;
- Uncertainties assigned to reference values.

Generally speaking, the laboratory should choose between general values (typically requirements of the test method) and values that represent the actual life of the laboratory.

For example, if the temperature is a source of uncertainty, the interval of tolerance of the standard [15°C;25°C] or the actual interval of performance of test within the lab [18°C;23°C] may be chosen. The first option is easier, the second requests mor work but is better for the lab.

Evaluating the effect of individual source of uncertainty on the final result:

The transformation of the scatter on an input parameter needs to be translated into a scatter in the final result. That is to say, it is needed to determine how much the final result is changed by the small changes that actually occur for the input parameter. The corresponding ratio is sometimes called coefficient of sensitivity. This can be achieved by the following methods:

- When the input parameter is a part of the variables of an equation that produces the final result, partial derivation of this equation with respect to the input parameter. GUM [2] provides the derivative of the usual functions (typically, addition, multiplication, polynomials, square roots). These derivative functions can then be used with relevant interval of scatter of input parameters. These intervals of scatter may be those of the standard or those that actually occur within the lab (which can be narrower). When this operation is complicated (because the initial equation is complicated), these coefficients of sensitivity can be determined by the Monte-Carlo method;

  Example 1: The linear mass LM *of a concrete reinforcing steel is determined on a piece of it which mass* M *and length* L *are measured.* LM *is then given by the equation* $LM = M/L$*. In that case, it is easy to find that* $u_M = \Delta M/M$ *and* $u_L = \Delta L/L$*, where ΔM and ΔL represent the scatter encountered in the lab around the "true" value of M and L respectively.*

  Example 2: ISO TR 15263 provides the results of derivation for some complicated formulas concerning tensile test of metals.

- When the input parameter is a numerical variable which is not part of an equation that produces the final result, it is needed to determine the effect of a controlled variation of this parameter on the final test result. It can be achieved by a GUM [2] method A, by internal experience of the lab or from an external source;

  For example, in many test methods, the temperature of performing the tests needs to be controlled. In those cases, it is needed to determine how much the final test result is changed from variations of temperature actually encountered in the laboratory.

- When the input parameter is a qualitative variable, it is needed to determine the effect of a controlled variation of this variable on the final test result. It can be achieved by a GUM [2] method A partial experiment, by internal experience of the lab or from an external source;

  Example 1: The shear force of a knot of a welded fabrics is determined with a tensile test for which the test specimen is hold in a "test specimen holder". The reference standard ISO 15630-2 provides several qualitative requirements for this test specimen holder like "prevent the transverse wire to rotate". It is then needed to determine to which extent the design of test specimen holder used in the lab produces deviations in the shear forces determined by it.

  Example 2: Effect of the material is a quasi-qualitative variable, which is quite complicated to deal with when no GUM [2] method A partial experiment is achieved.

- Some test methods produce a curve linking 2 or more parameters. The test result can be the curve itself or parameters that are computed from it. In those cases, it is needed to determine the effect of a controlled variation of this variable on the curve or on the parameters that are computed from it;

  Example 1: Infrared transmittance or reflectance curves, that are used to identify a product by comparison to reference curves of known products.

  Example 2: Tensile test of metals against ISO 6892-1, which raw results are expressed as a force-elongation curve. This standard defines several parameters and how to determine them from the curve. The determination of the $R_{p0,2}$ *requests a processing of the curve that is neither fully quantitative not fully qualitative.*

When the impact is determined as an overall interval, this overall interval should be divided by 4 (corresponding to an IC95% for Gaussian distributions) to be consistent with *u* which is a standard uncertainty.

For example, if the temperature is a source of uncertainty and the interval of tolerance on it set up at [15°C;25°C], the impact of a variation of (25-10)/4 = 2,5°C should be taken into account.

In some other cases, other types of distributions may apply. GUM [2] provides the equations to compute *u* for a selection of types of distributions encountered in practice. The monte-Carlo method can be used to determine *u* values for types of distributions for which no information is available.

For example, u *values related to the distribution of the resolution of the test equipment is usually a square one. In that case,* $u = a/\sqrt{3}$*, where a is the interval between two graduations of a measuring instrument.*

If the job has been conducted properly, all the u_i must be expressed in the unit of the test result or in percentage of it. If this condition is not fulfilled, the determination of *U* is likely to be wrong.

Computation the combination of all effects:

This computation is achieved by using Equation (8).

9.3 Conclusions about GUM method B

Because it does not request any experiment, probably some 90% of determinations of uncertainties are made using this method. It also has basic reasons to lead to underestimated uncertainties, and that is why most of uncertainties claimed by laboratories are underestimated. These reasons are detailed here after:

1. The bias is regarded as one of the random causes of uncertainty. This leads to underestimation of *U*, see § 10;
2. The unknown contributions are not taken into account;
3. This method is used because it avoids experimenting. Consequently, even when some experiment would be absolutely necessary (i.e. almost each time when a qualitative factor occurs), it is not in practice. We found

out many times during our experience of accreditation auditor that qualitative input parameters are systematically regarded as having a negligeable influence, just because considering them seriously is much complicated. Consequently, the main contributions are often omitted, what leads to underestimation, see § 10.

10 Issues and tools that apply to several of the described methods

10.1 Non numerical formats for test results

10.1.1 Introduction

Test results may be expressed in non-numerical formats like:

- Classification into categories (for example sweetness and aromas of wine), and whether these categories can be ordered or not (sweetness can be ordered from “not sweet” to “very sweet”, while aromas cannot);
- Binary results (for example pass/fail or presence/absence of a polluting agent) that can be regarded as a classification into 2 categories;
- Category percentages (for example classification of graphite in cast iron against ISO 945, where test results are provided in % of categories defined by standard pictures displayed in the standard);
- Curves (typically infrared spectrometry in which the IR curve is compared to a reference to identify which product the test item belongs to).

All tools and methods usually described to determine uncertainties cannot be implemented to such formats. Two possibilities can be implemented to overcome this difficulty:

1. Transform the non-numerical results into numerical ones;
2. Use statistics that can be implemented on non-numerical information (i.e. non-parametric statistics).

This issue is also extensively dealt with in another context in ISO 33407 [14].

10.1.2 Test results expressed as categories

Statistics applied on values of categories:

It is incorrect to calculate the mean values and standard deviations of category values, even when they are ordered, because category values are not real "numerical values" but "names" which can be transformed into any other type of label (letters, symbols, etc.) without any loss of information.

However, some test methods define categories in a such a way that they are actually “numerical values”.

Example: ISO 643 (determination of the grain size of metals) and ISO 4967 (determination of non-metallic inclusions in steel) define categories (of grain size, size of non-metallic inclusions respectively) as ranges of sizes that are regularly distributed, following a power law equation. They are then a kind of “numerical value” rounded in a non-decimal way.

The first step is then to check how the reference document for the test method defines the categories. If they are defined as regularly distributed ranges of numerical values, usual numerical techniques can be applied.

Use of underlying numerical values

In some cases, test results that expressed as categories are based on underlying numerical values that cannot be regarded as regularly distributed but can be determined by other testing methods. Usual methods can then not be applied to rank of them.

Example: Grade of pollution by a polluting agent. In that case, another testing method may provide numerical values of presence (for example in g/cm³) of this polluting agent.

In that case, it is possible to draw curves of probabilities of each test result as function of the underlying numerical value (see Figure 4) and deduce uncertainties about test results expressed as categories in the same format (i.e. probabilities of being in a given category). However, this requests to determine those curves of probabilities, what represents a big amount of work.

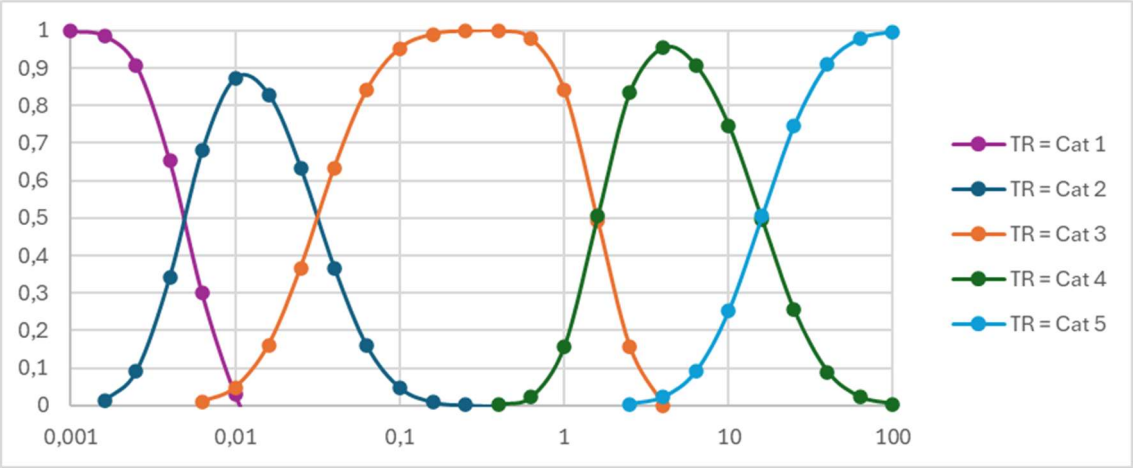


Figure 4: Probability for a sample to be classified in a category among 5 ordered ones, as function of the underlying numerical value *V*.
Example: if *V* = 10, 75% of test results (*TR*) are *Cat4* and 25% are *Cat5*.

Use of indexes based on ranks of categories:

ISO 13528 [3] provides information on how to use Gower indexes for transforming a set of results expressed as ordered categories into a set of results expressed as numerical values. These transformations are based on equations like Equation (9) (some other equations of similar form may be used), as follows:

$$NTR = \frac{i - \bar{\imath}}{N} \tag{9}$$

Where NTR *is the test result transformed into a numerical format,*
i *is the rank of the category*
$\bar{\imath}$ *is the average rank of the test results, usually computed as the rank of the median test result,*
N *is the number of categories into which the test results are distributed.*

Example: If a set of test results is [1,5 – 2 – 2 – 2,5 – 2,5 – 2,5 – 3 – 4] among categories that can be [0 – 0,5 – 1 – 1,5 – 2 – 2,5 – 3 – 3,5 – 4 – 4,5 – 5], that is to say 11 different possibilities, then $\bar{\imath} = 6$ *(rank of 2,5 among the possible test results), N = 11,* NTR *are respectively* $NTR = (4-6)/11 = -0{,}182$ *for TR = 1,5,* $NTR = (5-6)/11 = -0{,}091$ *for TR = 2, NTR = 0 for TR = 2,5, NTR = 0,091 for TR = 3 and NTR = 0,273 for TR = 4.*

By construction, Equation (9) produces NTR results belonging to the interval]-1;+1[. ISO 13528 [3] states that parametric statistics can be applied to those types of values. Note that the limits computed to define uncertainties

is generally somewhere between two categories. They need then to be transformed back into categories expressed as percentage of categories, as shown in the example below.

Example: If the set of test results [1,5 – 2 – 2 – 2,5 – 2,5 – 2,5 – 3 – 4] is the result of a partial GUM method A experiment as exposed in § 8.3, the mean value is then 0, the standard deviation is 0,137. Back to categories, the mean value is $11x0 + 2{,}5 = 2{,}5$*, the standard deviation is* $0{,}137x11 = 1{,}51$*, i.e. 151% of a category, lower limit for the IC95% is category of rank 6 minus 2x151% of a category, i.e. 2% of category of rank 2 and 98% of category of rank 3, i.e. not more than 2% of results "0,5" and no result "0". In the same the upper limit of IC95% is not more than 2% of results "4,5" and no result "5".*

When the categories are not ordered, ISO 13528 [3] proposes to order them from the most frequent to the less frequent and compute the Gower indexes as describes here upper.

Example: If a set of test results is [A – L – L – F – F – F – K] among categories that can be [A – B – D – K – L – M], that is to say 6 different possibilities, then category F stands for rank 1, category L stands for rank 2, categories A and K stand for rank 3 and other categories stand for rank 4. The average value is category "L" of rank 2.

It is then possible to apply Equation (9) to these test results.

Example: If a set of test results [A – L – L – F – F – F – K] is the result of a partial GUM method A experiment as exposed in § 8.3, the mean value is then -0,024, i.e. 86% of results "L" and 14% of results "F", and standard deviation is 0,154, i.e.0,154 x 6 = 93% of a category. In the same way than before, IC95% contains categories "F" and "L", not more than 86% of "A" and "K" results together and no results "B" or "M".

It shall be noted that these Gower indexes are computations on ranks of categories, and that other non-parametric methods also based on ranks may be used. A large range of ISO standards deal with these non-parametric statistical methods.

Statistics applying to proportions:

A large set of statistical methods were developed to compute intervals of confidence of proportions, particularly for polls. These statistics can seldom be used because they normally request a large quantity of test results. But when a large amount of test results is available, these methods provide better results than Gower indexes. Information about these methods can be found in ISO standards (particularly ISO 16269-6 [15]).

10.1.3 Binary results

Binary results can be regarded as a special case of categories with a number of categories *N* = 2. Proposals of § 10.1.2 can then be applied.

Another way to deal with binary results is to use the Binomial statistical law to determine the probabilities associated to each test result. Derived from the Binomial law, a large amount of scientific literature and of standards exists on the issue of false decisions (risks of wrong acceptances and risks of wrong rejection).

In the same way than upper, when a large amount of test results is available, methods dealing with IC on proportions are recommended.

10.1.4 Results expressed as percentages of categories

In these methods, results are expressed as percentages of categories.

Example: The test result is expressed as "A": 25%, "B": 40%, "C": 35%.

In those cases, the TR needs to be translated into 2 figures:

- One expressing the central value;

- One expressing the usual scatter that can be expected for one TR.

  Example: If the test result is expressed as "A": 25%, "B": 40%, "C" : 35% from ordered categories (A<B<C), the central value may be computed as the median (i.e. "B") or as a mean value using the Gower indexes (i.e. 93% "B" and 7% "C"). The scatter may be computed from the standard deviation (i.e. 0,301 or 30 % of a category).

Usual techniques concerning mean values and standard deviations may then be used to compute IC.

10.1.5 Results expressed as curves

In most of these cases, the reference documents that describe the test methods provide comprehensive requirements on how to determine numerical values from the curves.

Example: Tensile tests basically consist in plotting the diagram force – elongation that is recorded during the test. ISO 6892-1 (tensile tests on metals) define afterwards how typical parameters (Yield strength, Ultimate tensile strength, Elongation at maximum force, Elongation at rupture, …) shall be determined from this diagram.

But it happens that the curve itself stands for test result. In those cases, in order to determine the related uncertainties, the laboratory may:

- Define numerical parameters that characterises the diagram;

  Example: On IR spectrometry diagrams used for identifying paints or plastics, determine the wavenumber of the 10 main peaks).

- Define non numerical parameters that characterises the diagram;

  Example: On IR spectrometry diagrams used for identifying paints or plastics, determine the intensity of the 10 main peaks, qualified as "Important", "Medium", "Low". It is also possible to rank them from the most intense to the less intense.

- When the diagram is intended to be compared to a reference diagram, consider this as a binary test result.

  Example: For IR spectrometry diagrams used for identifying paints or plastics, determine the risk of wrong decision as described in § 10.1.3.

10.2 Change of variables

Some of the equations used to determine uncertainties (typically Equation (8)) do not request normality for the distribution of tests results, while some other do (typically an IC is determined, be using an enlarging coefficient, be when determining an IC on a SD as in § 10.7). Even more, the determination of SD can be significantly affected when a significant asymmetry occurs.

In any cases, determinations are of better reliability when distributions of data are at least symmetrical. Typical situations where this does not happen are when the scatter of results is large and:

- A physical or technical limit occurs for test results;
- Test results are a proportion, which implies two technical limits, i.e. 0% and 100%.

Case where one technical limit occurs:

In many cases, test results cannot be negative for technical reasons. Sometimes, other limits apply (for example, in tensile tests, from their definitions, the ratio R_m/R_e cannot be lower than 1 and A_{gt} cannot be lower than R_m/E). Limits can actually be minimum as well as maximum. In those cases:

- When the probability that some test results are close to the limit is low, the distribution of test results is not significantly altered, and normal distributions of test results can be assumed;

- Otherwise, an asymmetry is likely to happen, that may significantly affect the determinations of uncertainties. When 0 is the lower limit for test results, this problem should be considered when *CoV* (coefficient of variation) is less than 0,15.

A transformation of test results *TR* into *tr* = log(*TR*) is a very effective way to cope with an asymmetry problem. Calculations are then operated on *tr* values and results need to be transformed back into the initial scale afterwards. This can be achieved using the back transformation $TR = 10^{tr}$ (when logarithms of base 10 are used). This avoids the absurd situations where lower limits of IC are impossible values.

Example of an asymmetrical distribution which mean value is μ *= 100 and standard deviation is* σ *= 40 (*CoV *= 0,4). After log transformation, the mean value* μ *= 2 and the standard deviation is* σ *= 0,2 (better estimated because less affected by outliers of the upper side of the distribution). IC95% is then* $[10^{2-2.0,2};10^{2-2.0,2}]$*, i.e. [40;250] to compare with [20;180] without transformation.*

Cases where 0% and 100% limits occur:

Same problem may occur in some cases of proportions for which the scatter is important or when the uncertainty depends on the level of the test result.

In those cases, transformations using Equations $tr = \log(TR/(1-TR))$ (or $tr = \log(TR/(100-TR))$ if *TR* is expressed as a percentage) or $tr = (TR-0{,}5)/\sqrt{TR.(1-TR)}$ (or $tr = (TR-50)/\sqrt{TR.(100-TR)}$ if *TR* is expressed as a percentage) can be effective to:

- Transform asymmetrical distributions into symmetrical ones;
- Make *u* independent from the level of the measurand.

Example: The determination of the granulometry of a cement requests to determine the curve of mass percentages of particles as function of their size. In practice, this is achieved for a series of sizes between 1,25 µm to 100 µm which distribution follows the recommendations of ISO 3 [13]. Using one of the upper proposed transformations enables to:

- *Make* u *independent from the particle size (the actual value is usually not far from 4%);*
- *Increase the number of test results available to determine uncertainties;*
- *Determine reliable uncertainties for the tails of the distribution, i.e. for lower and upper sizes of particles.*

10.3 Issues linked to resolution and rounding of test results

It is reminded that resolution is the minimum step that can happen between 2 measurements or test results. Its effect, as well as rounding, is in most cases to reduce the estimation of the related standard deviation.

Example: If a set of 7 estimates of a test result which true value is 3,141592653578789… and which true standard deviation of estimation is 0,01, the standard deviation of estimates rounded to the closest 1 (r = 1), closest 0,1 (r = 0,1), closest 0,01 (r = 0,01) and closest 0,001 (r = 0,001) distribute as shown in Figure 5.

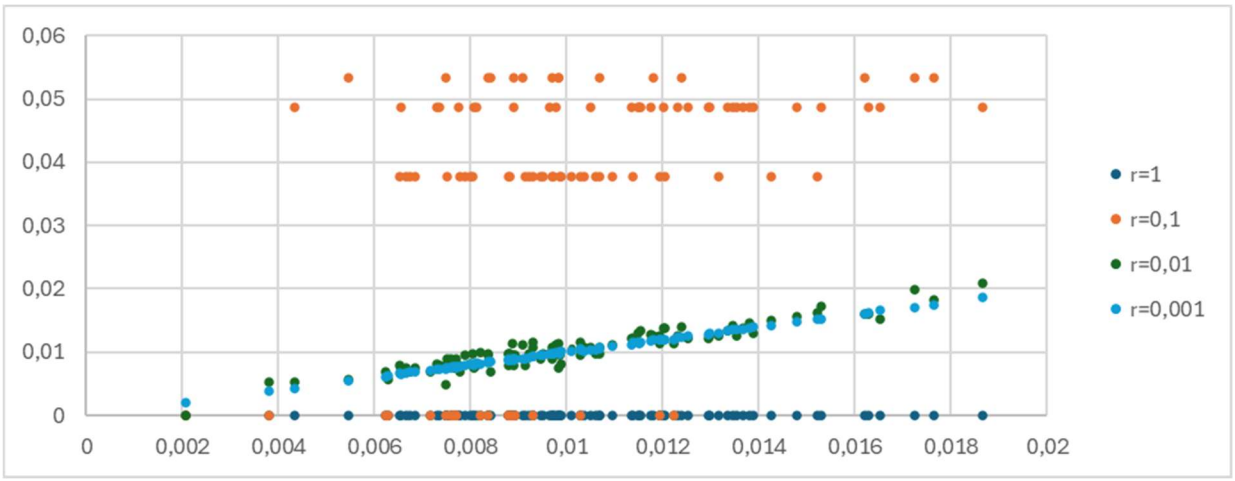


Figure 5: Distributions of s/s_{true} ratios (ratios of s computed from rounded values against s computed from not rounded values)

It can be seen from this example that:

- *When* r *= 1 (100 times the true standard deviation of estimation), the standard deviation of estimation is always reduced to 0;*
- *When* r *= 0,1 (10 times the true standard deviation of estimation), the standard deviations of estimations distribute on a very small number of occurrences (i.e., in this case, 0 – 3,8 – 4,9 – 5,3 times the true standard deviation), that is to say, a very poor quality of estimation;*
- *When* r *= 0,01 (1 time the true standard deviation of estimation), the IC95% of the standard deviation of estimation is]0,8.s;1,34.s[, that is to say an acceptable but significantly affected estimation;*
- *When* r *= 0,001 (0,1 time the true standard deviation of estimation), the IC95% of the standard deviation of estimation is]0,97.s;1,03.s[, that is to say, the estimation of the standard deviation is not affected by the rounding or resolution.*

This confirms the usual rule saying that the resolution / rounding should be at least 1/10 of the actual standard deviation. This is also consistent with conclusions of § 10.8.

It should be noted that when r is in the same range than the standard deviation of estimation, the behaviour of the s/s_{true} ratios depends on the number of estimates included in the set and on where the unrounded value is located in the step.

In the example here upper, when r *= 0,01 (1 time the true standard deviation of estimation), the standard deviations of estimations distribute on 4 different values, corresponding to cases where:*

1. *All estimates are rounded to the same value;*
2. *One estimate is rounded to one side of the step and six on the other side;*
3. *Two estimates are rounded to one side of the step and five on the other side;*
4. *Three estimates are rounded to one side of the step and four on the other side.*

that is to say, n*/2+1 or (*n*+1)/2 occurrences, depending on whether* n *(number of estimates included in the set) is an even or an odd number.*

Moreover, the actual values of s/s_{true} *(i.e., in this example, 0 – 3,8 – 4,9 – 5,3 times the true standard deviation), depend on where the unrounded value is located in the step (i.e. 0,41592653578789… times the value of the step in the example). If the true value had been closed to a rounded one, the actual values of* s/s_{true} *would have been close to 0.*

It should be also noted that the level of rounding / resolution may be hidden by a subsequent calculation. Consequently, care should be taken to track such possibilities.

Example: For the determination of strengths during tensile tests on metals against ISO 6892-1, the resolution of the test result is mainly governed by the resolution of the force sensor. If the range of the sensor is 628 kN and the number of increments is 24575, the measured force is rounded to the closest 25,5544 N, which stands then as the resolution. On the other hand, the section is measured with dimensions rounded to the closest 0,005 mm, which generates also a hidden resolution for the measurement of the section (for example 0,03142 mm²). The test result (division of the 2 measurands) gets then also a resolution, which can be irregular.

When such a ratio is not reached, the resolution / rounding should be ameliorated. This can be achieved by:

- Increase the basis of measurement;
- Improve the accuracy of the measurement device.

However, in some cases, because of technical reasons, it is not possible to increase the resolution. In those cases, a contribution of the resolution should be added to the uncertainty. GUM [2] provides the necessary guidance to achieve it.

In most cases, it consists in adding a contribution $u = a/\sqrt{3}$, where a is the increment. This is valid when the true value is randomly distributed between the 2 limits of the increment (i.e. square distribution).

10.4 Cases where several values of uncertainties are available for several different situations

It may happen that several estimations of uncertainties from different sources are available in the laboratory, each of them is related to different sets of conditions (material, test equipment, operators, …).

For example, this situation may occur when the lab is conducting several control charts, each of them related to a testing machine.

In such cases, the lab has got 2 possibilities:

1. Compute as many values of *U* as the number of available sets of data. A more detailed information is then available for the lab, but it then needs to manage a greater number of data when, for example, a customer requests the *U* value attached to its test results;
2. Consolidate the *U* values obtained from each set of test results into a general *U* value, valid for all test results that are produced by the lab. In that case, a weighted quadratic mean value of *U* should be used, as shown in Equation (10).

$$U = \left|\frac{c_1.B_1 + \cdots + c_i.B_i + \cdots + c_n.B_n}{c_1 + \cdots + c_i + \cdots + c_n}\right| + k.\sqrt{\frac{c_1.u_1^2 + \cdots + c_i.u_i^2 + \cdots + c_n.u_n^2}{c_1 + \cdots + c_i + \cdots + c_n}} \quad (10)$$

where B_i *is the bias (i.e.* $m_i - X_{RM}$*) from the* i^{th} *set of data*
u_i *is the standard uncertainty from the* i^{th} *set of data,*
c_i *is a weighting coefficient for the* i^{th} *set of data,*
n *is the number of sets of data used to determine the* u_i*,*
k *is the enlarging coefficient.*

Note that:

- The sign of each B_i should be kept in the intermediate calculations and the absolute value should be applied on the sum of all positive and negative values of B_i;
- The c_i coefficients are intended to make the global *U* value more representative of the whole quantity of uncertainties that applies to the overall production of test results of the lab. They should then be chosen as a function of the proportion of test results coming from each of the situations represented by each of the u_i

individual standard uncertainties. It is possible to manage c_i coefficients so that their sum is equal to 1: this simplifies Equation (10) and shows more clearly which situations are the most contributing to uncertainties.

10.5 General issues concerning estimation

Rules for estimation applies when the value of a parameter is computed from a selection of occurrences of a parameter.

Example: when a mean value of a population is computed from 10 values, all of them belonging to this population.

It is most often the case in the determination of uncertainties, but not always. When the totality of possible occurrences is included in the experiment the considered parameter is then not anymore estimated, it is computed. In particular, the formula for determining the standard deviations is not the same in the two cases: Equation (11) when all occurrences are taken into account in the calculation and Equation (12) when a selection of occurrences are taken into account in the calculation.

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \tag{11}$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{12}$$

where σ *is the standard deviation to compute,*
s *is the estimate of a standard deviation σ to compute*
x_i *are the individual values of the set used to compute the standard deviation,*
$\bar{x}$ *is the mean value of the* x_i*,*
and n *is the number of values used for computing the mean value.*

Example 1: When a partial GUM method A experiment is organised to determine the contribution of operators and this experiment includes all operators of the lab, and each mean value of operators is supposed to be accurately known, Equation (11) applies and the value of the standard deviation is known exactly. Consequently, statements of § 10.7 do not apply to this case.

Example 2: In the same experiment, each operator could potentially repeat the tests an infinite number of times. The set of value used to compute the repeatability standard deviation is then a selection of all possible performances. Consequently, Equation (12) *applies and the value of the standard deviation is not known exactly. The statements of § 10.7 can be used to determine the IC95% of the true σ value.*

10.6 Issues concerning the estimation of a mean value

The distribution law of estimates of mean values is as described in Equation (13) (origin: ISO 2854 [16]).

$$m \approx N(\mu, \frac{\sigma}{\sqrt{n}}) \tag{13}$$

where m *is the estimate of a mean value,*
μ *is the mean value to be estimated*
σ *is the standard deviation of the corresponding population,*
and n *is the number of values used for computing the mean value.*

When both μ and σ are unknown, the Student tables can be used to take into account the uncertainty related to the estimations of these parameters.

10.7 Issues concerning the estimation of a standard deviation

The distribution law of estimates of variances is as described in Equation (14) (origin: ISO 2854 [16]).

$$(n-1)\,\frac{s^2}{\sigma^2} \approx \chi^2_{n-1} \tag{14}$$

where s is the estimate of a standard deviation,
σ is the standard deviation to be estimated,
and n is the number of values used for computing the standard deviation.

The distribution law of estimates of standard deviations can then be deduced by algebraic transformation of Equation (14) as stated in Equation (15).

$$s = \sigma \sqrt{\frac{\chi^2_{n-1}(P)}{n-1}} \tag{15}$$

where s is the estimate of a standard deviation,
σ is the standard deviation to be estimated,
P is the corresponding theoretical cumulated probability,
and n is the number of values used for computing the standard deviation.

In practice, *s* is known (determined from the experiments) and the information of interest is the value of *σ*. We can then compute limits for an IC95% by computing the *σ*/*s* ratios for *P* = 0,025 and *P* = 0,975. These limits are displayed in Table 3 for increasing values of *v* = *n*-1 (number of degrees of freedom).

Table 3. Lower and upper limits of the IC95% interval of σ/s ratios as function of v = n-1.

2	**3**	**4**	**5**	**6**	**8**	**10**	**13**	**16**	**20**
0,521	0,566	0,599	0,624	0,644	0,675	0,699	0,725	0,745	0,765
6,285	3,729	2,874	2,453	2,202	1,916	1,755	1,611	1,522	1,444
25	**32**	**40**	**50**	**63**	**80**	**100**	**125**	**160**	**200**
0,784	0,804	0,821	0,837	0,852	0,866	0,879	0,89	0,901	0,911
1,38	1,323	1,28	1,243	1,211	1,183	1,161	1,141	1,123	1,109

It can be seen that a large number of test results is needed to get an acceptable estimation of *σ* from available values of *s*.

When the standard deviation is estimated from a series of repetitions (for example in Equation (7)), the same table applies with *n*-1 replaced with *n*.(*r*-1), where *n* is the number of series and *r* is the number of repetitions.

For example, in an experiment as described in § 7.3, if 50 configurations are tested with one repetition for each, the estimate of repeatability computed with Equation (7) has got an IC95% as stated in Table 3 with v = 50.(2-1) = 50, i.e. [0,837;1,243].

This increase in *v* values should be considered when such experiments are designed.

10.8 Issues concerning the combination of standard deviations

In many cases, a combination of standard deviations is applied (see Equations (3), (4), (5), (6), and (8)). It shall be noted that these equations of combination of standard deviation is valid even when the corresponding distributions are not Gaussian.

It must be however noted that the standard deviations included in the formula must be independent. If not, a term of correlation between them should be introduced.

On a mathematical point of view, a correlation is proved when *r* (coefficient of correlation) is significantly different from 0, but the reciprocal is not true (i.e. a real correlation may exist even if *r* is not significantly different from 0).

> *This latter situation is particularly likely to occur when the number of pairs introduced into the correlation is insufficient and/or when the correlation is non-linear (for example, a hypothetical correlation in $x^2+y^2=1$ would produce r coefficients always close to 0).*

In practice, when a correlation has technical reasons to be present, it is usually easier to determine a global *u* value that encompasses both effects and their correlation.

> *For example, if $u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$, but if u_1 and u_2 are in fact positively correlated (r = 1), the equation must be transformed into $u_{total} = \sqrt{(1 + 0{,}5)^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$, and if u_1 and u_2 are in fact negatively correlated (r = -1), the equation must be transformed into $u_{total} = \sqrt{(1 - 0{,}5)^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$. The corresponding u_{total} results are then respectively 0,55 and 1,517 instead of 1,141.*

In most cases, the correlations are never complete and appropriate equations as a function of *r* must be used. When only two variables are correlated with a correlation coefficient *r*, the overall variance of these two variables is $\sigma^2 = \sigma_1^2 + 2.r.\sigma_1.\sigma_2 + \sigma_2^2$.

The main contributions introduced in the formula mainly govern the final result, as shown in the example here after, where contributions are classified from the most important to the less important.

$u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2} = 1{,}14127$, but when the less important contribution is omitted,

$u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2} = 1{,}14018$,

$u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2} = 1{,}13578$,

$u_{total} = \sqrt{1^2 + 0{,}5^2} = 1{,}11803$,

$u_{total} = \sqrt{1^2} = 1$

We can see that only the 2 first terms are of importance.

However, with respect to § 10.7 here upper, uncertainty on estimations of *u* is usually in the range of 1 to 4 (when it is computed from 6 test results or less), and the mutual real importance of the two first terms are uncertain. A Monte-Carlo experiment was performed to cope with this issue. This experiment was conducted on an example where $u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$, each of these u_i are determined as standard deviations determined from *n* results. Resulting σ/s ratios as function of *n* (number of repetitions used to compute *s*) and N_t (number of terms taken in u_{total} of here upper) are displayed in Figure 6. Central curves are averages, upper and lower curves are limits of IC95%. Detailed results are available in § 10.11.

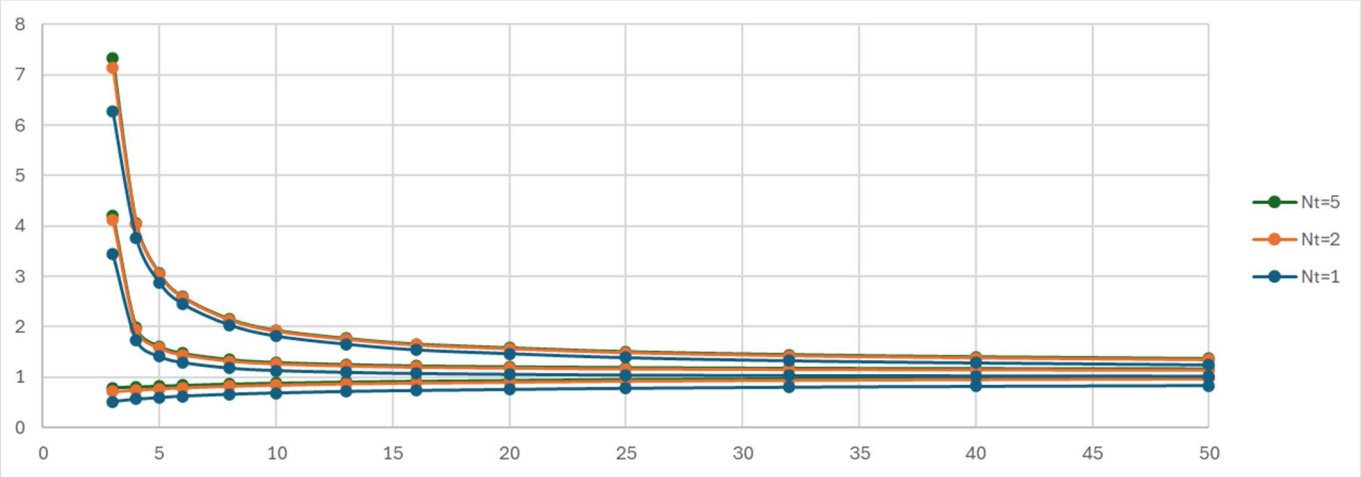


Figure 6: *σ/s* ratios as function of *n* (number of repetitions used to compute *s*) and N_t (number of terms taken in u_{total} of here upper), Central curves are averages, upper and lower curves are limits of IC95%.

It should be noted that, in real life, rounding effects tend to reduce the estimated *s* value, i.e. to increase *σ/s* ratios, what amplifies the phenomena that are observed in here.

It can be seen from this experiment that:

- Only the 2 first terms are of importance (curves N_t = 2 and N_t = 5 are almost coincide);
- The distances between average curves and IC95% limit curves is far greater than the distances between curves N_t = 1 and N_t = 2. However, unprovided results show that the centile 97,5% of the ratio of 2nd term to 1st term σ_2/σ_1 is more than 1 when *n* < 10. It is then advisable to determine accurately all terms that are at least 50% of the main term (here the 2 main terms);
- Because of the strong asymmetry of the distributions of *σ/s* ratios, the average values of them is significantly greater than 1, even for large values of n. This appears to be another major cause of underestimating *u* values.

For example, when the determination of u_{total} *from 5 terms, each of them coming from computations of SD from 3 test results, lead to* u *= 1,413, the average true value of it is in fact 4,21 and the limits of IC95% for it are 0,80 and 7,93.*

Consequently, when the determination of U requests a lot of experiments, a procedure in 2 steps can be useful, as follows:

1. Make a preliminary study of the contributions of each source of uncertainty;
2. Make a detailed study of the contributions of the sources that are more than 50% of the main one.

10.9 Impact of choice of classification of a source of uncertainty as bias or random error

Impact on *u* and *U* of classification of a bias as a random error:

When a bias is determined and computed as a bias, the *U* value is computed with Equation (3), (4) or (6). When it is regarded as a random error, the *U* value is computed with Equation (8). When *e* << *B*, this leads to significant difference in the calculation of *U* (ratio tends to the value of *k*), as shown here after (example with *B* = 1, *e* = 0,1 and *k* = 2):

- Option with *B* regarded as a bias: $U = 1 + 2 \times 0{,}1 = 1{,}2$;
- Option with *B* included in the random error: $U = 2.\sqrt{1^2 + 0{,}1^2} = 2{,}01$;

In the other cases, important differences may occur for *u*, but not for *U*, as shown here after (example with *B* = 1, *e* = 1 and *k* = 2):

- Option with *B* regarded as a bias: $u = 1$ and $U = 1 + 2 \times 1 = 3$;
- Option with *B* included in the random error: $u = \sqrt{1^2 + 1^2} = 1{,}4$ and $U = 2 \times 1{,}4 = 2{,}8$.

Impact on *IC95%* of classification of a bias as a random error:

As seen before, GUM [2] proposes to use an enlargement coefficient *k* to determine an IC, assuming that the overall distribution of test results is Gaussian, what is justified by the central limit theorem of statistics, and often correct on the practical point of view. *k* is chosen according to the with of the desired IC (for example, *k* = 2 for the IC95%).

However, this does not apply correctly when *e* << *B*. In those cases:

- The *U* value is badly determined if the bias is not determined separately (see here upper);
- The effect of random error is deleterious only on one side of its distribution (see Figure 3). The *k* value should then be chosen for unilateral *P* values.

It follows that:

- When *e* << *B*, *k* should be chosen equal to 1,64 (IC95% unilateral) so that, in the former example, true $U = 1 + 1{,}64 \times 0{,}1 = 1{,}16$ instead of 1,2;
- In the other cases, choosing *k* = 2 is correct, so that calculations here upper are not changed.

Conclusions:

When *e* << *B*, including the bias *B* into the random term *u* leads to overestimate *U* by the factor *k* of enlargement. In other cases, including the bias *B* into the random term *u* leads to underestimate *u* but this underestimation is more or less cancelled by the application of the enlargement coefficient when the enlarged uncertainty *U* is computed.

10.10 Test results for given environmental conditions (typically, for given temperatures) or, more generally, as function of external conditions

It happens frequently that test results are needed for environmental conditions that are not the ambient ones. In those cases, the laboratory may choose among 2 options:

- Include the contribution of deviation from requested environmental conditions into the global determination of *u*;
- Provide two separate values of *u*, one for the deviation from requested environmental conditions and the other for the uncertainty on the test result itself, without including the impact of the deviation from the required environmental conditions.

The second option is easier, because it does not request to determine the impact of deviation of the environmental conditions on the test result (which can sometimes quite hard to do properly). Depending on the use of the test results, the first or the second option can be the best.

Example: Carbon steels may become brittle under a certain temperature called transition temperature, that is determined with a series of Charpy impact tests against ISO 148-1 as function of temperature.

When the test results are intended to build up this transition curve, it is better to express the uncertainties separately. Uncertainties can be represented as ellipses on the curves and their values are more reliable.

When the test results are used to declare conformity to a requirement (in the form KV_2 > 30J at -50°C) it is necessary to include the contribution of uncertainty on the testing temperature into the global one to make reliable the declaration of conformity.

As a conclusion, in such cases, the laboratory should choose its option in accordance with the intended use of the declared uncertainties.

10.11 The Monte-Carlo method

The Monte-Carlo methods are a large category of algorithms that use random numerical realisations of a given model. They are often used to solve mathematical or physical problems, difficult or impossible to solve by other methods. For a survey of the history and applications of the Monte-Carlo methods, see for example [17]. In the same way, JCGM 101 [18] from BIPM provides recommendations concerning the application of the Monte-Carlo method to the expression of uncertainties.

Hard calculations are needed to solve several of the issues of this document. In order to simplify these calculations, we used the Monte-Carlo method. In the frame of this study, determining centiles, medians or mean values of distributions request to solve some difficult integrals and to find zeros of polynomial equations. To avoid this, large series of random realisations are created, what enables to compute these centiles, medians or mean values.

However, using Monte-Carlo methods requests to use a model that represents reasonably well the situations that we want to deal with. To achieve this, an appropriate modelling is needed. In the frame of this study, these models are provided by Equations proposed along this document, in particular the fundamental Equations (2) and (3).

Using the Monte-Carlo methods also requests to use random input values. When several random values are necessary to produce one Monte-Carlo result and when correlations between them apply in the phenomenon to represent, these correlations must be incorporated in the input values of the computations. That can be a bit difficult to do properly.

To assure the validity of the conclusions, the random series need to be numerous enough, depending on many factors. A solution to control this is to divide these series into sub-groups. This enables us to compute the repeatability of the parameters that we are determining. This repeatability standard deviation is then used to determine an IC for each of the determinations, with an enlargement coefficient equal to 2.

Example: In § 10.8, a study of impact on the combination of the estimation of each of the u_i was conducted. This can hardly be performed by using mathematical calculations (it would have been completely impossible if the number n of tests used would have been different for each u_i, which is the usual situation). To cope with this issue, the Monte-Carlo method was used as follows:

- *For each u_i term of equation $u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$, a random potential true σ value was computed with Equation (15);*
- *Each computation of u_{total}, from $u_{total} = \sqrt{1^2}$ to $u_{total} = \sqrt{1^2 + 0{,}5^2 + 0{,}2^2 + 0{,}1^2 + 0{,}05^2}$ was computed;*
- *This was performed 200000 times in 20 subseries of 10000, for each value of n;*
- *For each of the subseries, the quadratic mean value QM and the centiles 2,5% and 97,5% were computed;*
- *The mean value and the standard deviation were computed from the subseries for the parameters (i.e. QM and centiles), enabling to compute an extended uncertainty 2.u on QM and centiles.*

In this example:

- *Equations (8) and (15) describe the behaviour of the phenomenon. We have a valid model, as requested to implement the Monte-Carlo method;*

- *Equation (8) requests independency of* u_i *to be valid. We can then validly use 5 random numbers for each Monte-Carlo realisation;*
- *The division of the 200000 realisations into 20 subgroups enable to compute 2.*u *and decide whether the number of realisations is enough or not. In the present case, all* u *are low enough for our application, except the average values for* n *= 3. If these values would have been needed for any use, the Monte-Carlo algorithm should probably be continued to get more accurate values.*

Results are provided in Table 4 as follows.

Table 4. Average, lower and upper limits of the IC95% interval of σ/s ratios as function of n *(number of repetitions) and* N_t *(number of terms in Equation of* u_{total}*).*

N_t	*n*	3	4	5	6	8	10	13	16	20	25	32	40	50
1	***Average***	3,44	1,732	1,414	1,292	1,183	1,133	1,098	1,074	1,058	1,044	1,033	1,027	1,021
	2u	0,26	0,011	0,005	0,002	0,001	0,001	0,001	0,001	0,001	0,000	0,000	0,000	0,000
	IC95%-	0,520	0,567	0,598	0,626	0,663	0,688	0,718	0,738	0,759	0,781	0,802	0,819	0,836
	2u	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,000	0,001	0,000	0,000
	IC95%+	6,272	3,761	2,877	2,455	2,038	1,819	1,658	1,544	1,468	1,389	1,328	1,284	1,246
	2u	0,041	0,016	0,012	0,006	0,004	0,003	0,002	0,002	0,005	0,001	0,001	0,001	0,001
2	***Average***	4,12	1,940	1,582	1,444	1,323	1,267	1,226	1,201	1,183	1,168	1,155	1,148	1,141
	2u	0,26	0,011	0,005	0,002	0,001	0,001	0,001	0,000	0,001	0,000	0,000	0,000	0,000
	IC95%-	0,714	0,748	0,773	0,792	0,823	0,845	0,870	0,887	0,904	0,923	0,941	0,956	0,970
	2u	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,000	0,001	0,000	0,000
	IC95%+	7,141	4,032	3,038	2,577	2,139	1,916	1,754	1,642	1,566	1,489	1,429	1,387	1,350
	2u	0,053	0,017	0,011	0,007	0,004	0,003	0,002	0,002	0,005	0,001	0,001	0,001	0,001
3	***Average***	4,19	1,970	1,608	1,467	1,344	1,287	1,246	1,220	1,202	1,186	1,174	1,166	1,160
	2u	0,26	0,011	0,005	0,002	0,001	0,001	0,001	0,000	0,001	0,000	0,000	0,000	0,000
	IC95%-	0,769	0,791	0,811	0,827	0,854	0,873	0,896	0,912	0,928	0,946	0,963	0,977	0,991
	2u	0,002	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,000	0,001	0,000	0,000
	IC95%+	7,287	4,056	3,056	2,591	2,152	1,929	1,768	1,656	1,580	1,504	1,444	1,402	1,366
	2u	0,055	0,016	0,012	0,007	0,004	0,003	0,002	0,002	0,005	0,001	0,001	0,001	0,001
4	***Average***	4,21	1,977	1,614	1,473	1,349	1,292	1,251	1,224	1,207	1,191	1,178	1,171	1,164
	2u	0,26	0,011	0,005	0,002	0,001	0,001	0,001	0,000	0,001	0,000	0,000	0,000	0,000
	IC95%-	0,789	0,804	0,822	0,836	0,862	0,880	0,903	0,918	0,934	0,951	0,969	0,983	0,996
	2u	0,002	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,000	0,001	0,000	0,000
	IC95%+	7,317	4,059	3,058	2,594	2,155	1,932	1,771	1,659	1,584	1,508	1,448	1,406	1,369
	2u	0,057	0,016	0,012	0,007	0,004	0,003	0,002	0,002	0,005	0,001	0,001	0,001	0,001
5	***Average***	4,21	1,979	1,615	1,474	1,350	1,293	1,252	1,226	1,208	1,192	1,179	1,172	1,165
	2u	0,26	0,011	0,005	0,002	0,001	0,001	0,001	0,000	0,001	0,000	0,000	0,000	0,000
	IC95%-	0,796	0,808	0,824	0,839	0,864	0,882	0,904	0,919	0,936	0,953	0,970	0,984	0,998
	2u	0,002	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,000	0,000	0,000	0,000

N_t	*n*	3	4	5	6	8	10	13	16	20	25	32	40	50
	IC95%+	7,322	4,060	3,059	2,595	2,156	1,933	1,772	1,660	1,585	1,509	1,449	1,407	1,370
	2u	0,057	0,016	0,012	0,007	0,004	0,003	0,002	0,002	0,005	0,001	0,001	0,001	0,001

10.12 Analysis of variances

Analysis of variances (ANOVA) are statistical techniques that are used to identify and quantify individual random effects in a measurement so that they may be properly taken into account when the uncertainty of the result of the measurement is evaluated.

Although ANOVA methods are applicable to a wide range of measurements, these methods are focused on standard deviations and then, cannot by themselves identify systematic effects that might be present.

GUM [2] Annex H provides a survey on how ANOVA techniques may be useful for the determination of uncertainties.

Understanding basics of ANOVA, in particular how degrees of freedom operate, helps to define more effective design of experiments by centring testing where it is needed.

11 Assessment of the quality of uncertainties determined by the laboratory

11.1 Introduction

Two different methods can be used to check the quality of the determination of uncertainties:

- The first one is based on one of the principles of ISO 21748 [11], that states that s_R (reproducibility of the test method) is a good basis for estimating an uncertainty;
- The second one is based on the statements of ISO 13528 [3], that proposes a ζ-score to assess uncertainties during PT exercises, which can be adapted for a use within a single laboratory.

Both options can be implemented by a single laboratory or by a PT provider for the participants of its PT programs.

11.2 Use of the reproducibility standard deviation

When the reproducibility standard deviation s_R can be known from external sources (typically from ILC or scientific documentation or standardisation), it can be taken as a basis for uncertainty.

In most cases, the laboratory is deemed to consider itself as an “average” one and consequently, which uncertainty is not significantly different from s_R.

In some cases, the laboratory may have good reasons to consider itself as a non-average one. Typically, it can be:

- A reference laboratory, that implement resources that are better than what is requested in the reference document for it (typically, test equipment more accurate than requested, environmental conditions narrower than requested (for example intervals of temperature of ±1°C compared to a requested ±5°C), etc. …);

For example, it can be a national laboratory that is intended to produce CRM.

- A routine laboratory, that implement resources that do not fulfil completely the requirements of the reference document but which practice is good enough for the intended use of the test results.

For example, a laboratory of a factory, intended to check the conformity of products that, from the knowledge of the manufacturing process, are very unlikely to be not fulfilling the product requirements.

In all cases, the ratio u_i/s_R should be consistent with what should be expected for the laboratory:

- $u_i/s_R < 1$ for reference laboratories;
- $u_i/s_R \cong 1$ for average laboratories;
- $u_i/s_R > 1$ for laboratories that do not fulfil completely the requirements.

11.3 Use of an adapted ζ-score

The method using ζ-scores as described in ISO 13528 [3] can be extended to assess the quality of uncertainties determined by the laboratory. For achieving that, Equation (1) can be reformulated into Equation (16), as follows:

$$\zeta = \frac{x_i - X_{RM}}{\sqrt{u_{RM}^2 + u_i^2}} \qquad (16)$$

where x_i *is any result obtained by the lab on a RM,*
X_{RM} *is the assigned value of the RM,*
u_{RM} *is the uncertainty on* X_{RM}*,*
and u_i *is the uncertainty claimed by the laboratory concerning this test result.*

As detailed in § 7.2, the RM does not need to be a CRM, but it needs to be a material which X_{RM} is known for another source than the tests performed within the laboratory (i.e. from an ILC, from other source of knowledge, from a collaboration of several laboratories, etc. …).

Of course, ζ-scores should not be computed from sets of test results that were used to compute u_i. Such a computation would be some kind of circular auto-confirmation.

By design, ζ-scores can be deemed to follow a Gaussian distribution which mean value is 0 and standard deviation is 1. Usual limits (i.e. 2 and 3) can then be used to check u_i values. Moreover, when a enough quantity of ζ-scores were computed from a series of x_i values, the standard deviation of ζ-values s_ζ can be computed, and this s_ζ should not deviate significantly from 1. Statements of § 10.7 can be used to check this.

When U includes a term of bias (in the form $U = B + k.u$), Equation (16) needs to be slightly adapted. The easiest way to do it is to use u_i values computed with the Equation $u_i = B/k + u$, so that both B and u represent properly their corresponding sources of uncertainty. In those cases, the only limit that make sense for ζ is k, i.e. 2 in most of cases.

Using Equation (16) on a single u_i value only enables to check whether is underestimated. When the u_i value is overestimated, the corresponding ζ-score becomes very small but no statistical test can put it in evidence.

On the contrary, the computation of s_ζ values enables to check both underestimation and overestimation of u_i values, but only at a level that implies several determinations of uncertainties:

- Overestimating u_i leads to s_ζ values significantly smaller than 1 and underestimating u_i leads to s_ζ values significantly greater than 1;
- s_ζ values make sense only for a set of several to many determinations of uncertainty.

The evaluation of the s_ζ values can therefore be used to assess whether there is a general problem in evaluating uncertainties, for example for a particular method of determining these uncertainties.

11.4 Results of assessment of uncertainties

Figure 7.a to i provide graphical representations of assessments of u_i performed in accordance with the proposals of the present § 11, from a selection of PT programs performed by CompaLab in 2023. In addition to these graphical representations, Table 5 provide the standard deviations of ζ-scores s_ζ for this selection.

It shall be noted that:

- In that exercise, s_ζ represent the standard deviations of ζ-scores of participants and not from a series of ζ-scores from assessment performed inside a same laboratory. Consequently, a significant deviation of s_ζ to 1 means something concerning the globality of participants rather than something concerning each individual one;
- Also because of this situation, some ζ-scores obviously appeared to be outliers that impacted much the calculation of s_ζ. These outliers were then suppressed from the data before the computations. The number of them for each program is mentioned in Table 5;
- In order to ease the reading, both Figure 7.a to i and Table 5 were ordered by increasing values of s_ζ.

Table 5. Standard deviations of ζ-scores s_ζ for a selection of 2023 PT programs

PT scheme	Chemical analysis– Low alloyed steel – Carbon content %	Tensile test– Tensile strength - R_m	Tensile test– Yield strength $R_{p0,2}$	Vickers hardness test– *HV10*	Charpy impact test – KV_2 energy	Charpy impact test – Lateral expansion	Shear force of welded fabrics	Tensile test– Elongation at rupture A_{5d}	Huey corrosion test – stainless steels – corrosion rate (48h)
Nb of participants	118	26	24	81	25	19	30	23	25
Nb of declared u_i	78	15	12	51	16	10	16	13	10
Percentage	66%	58%	50%	63%	64%	53%	53%	57%	40%
Nb of outliers for ζ-score	3	1	2	2	0	0	0	1	1
s_ζ	0,90	1,07	1,35	1,56	2,14	2,41	2,75	3,02	3,42

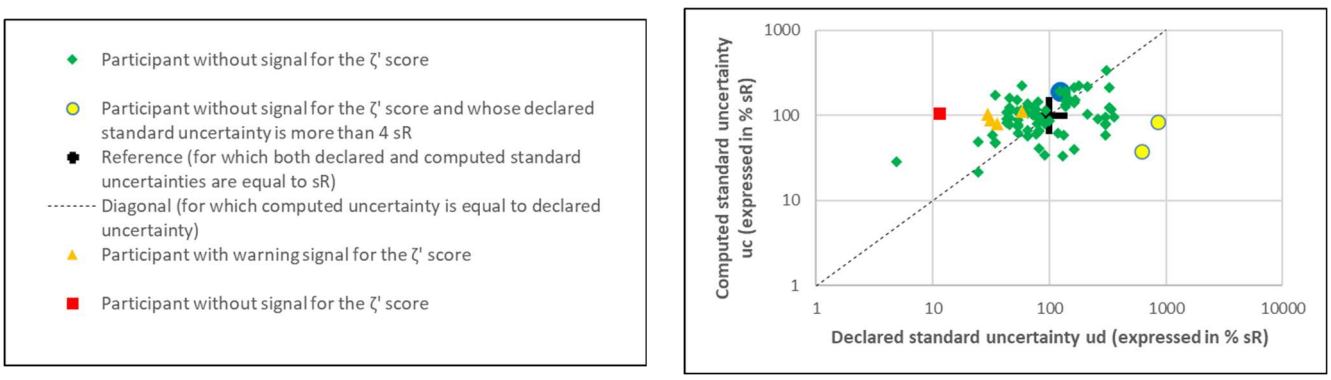


Figure 7.a: Results of 2023 of assessment of uncertainties for carbon content on low alloyed steel.

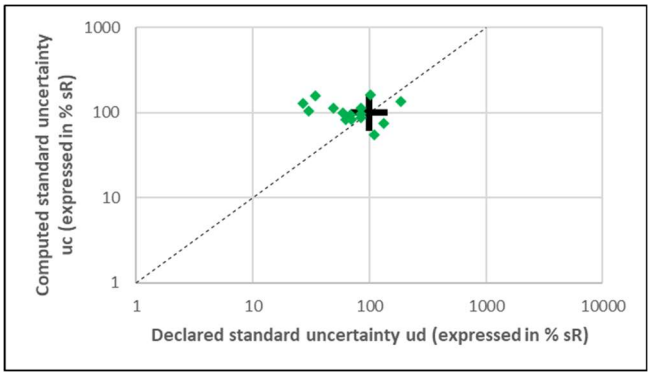


Figure 7.b: Results of 2023 of assessment of uncertainties for tensile strength R_m.

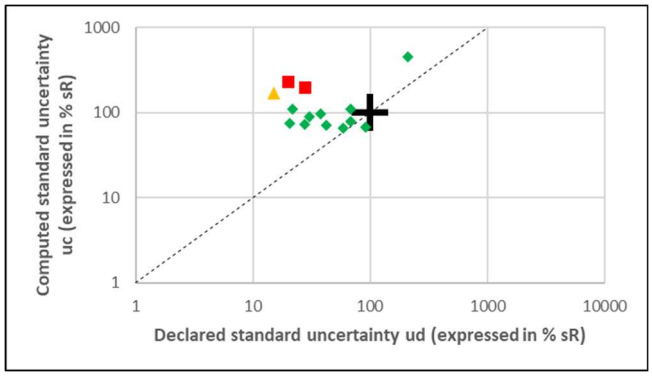


Figure 7.c: Results of 2023 of assessment of uncertainties for yield strength $R_{p0,2}$.

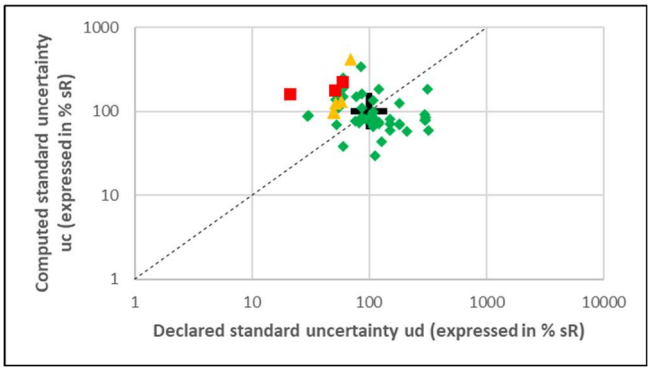


Figure 7.d: Results of 2023 of assessment of uncertainties for Vickers hardness *HV10*.

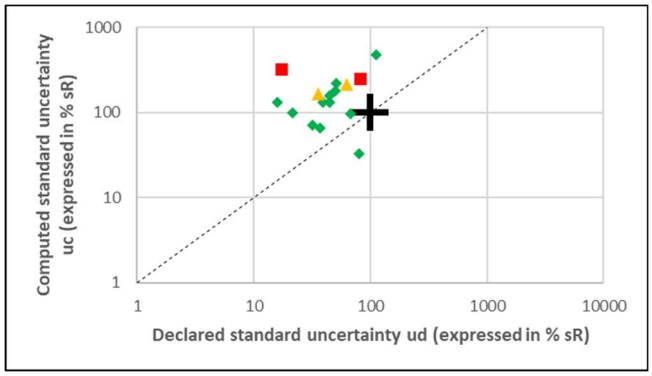


Figure 7.e: Results of 2023 of assessment of uncertainties for Charpy impact test, Energy KV_2.

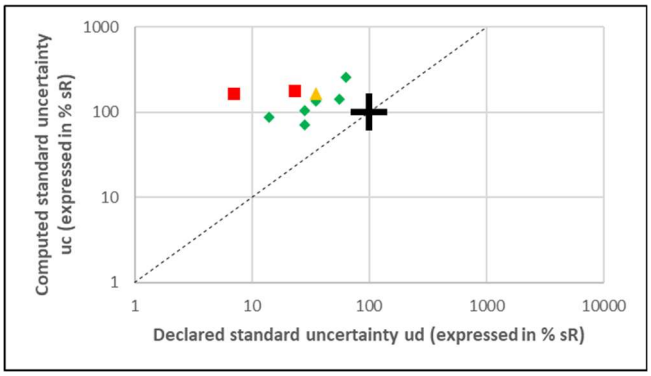


Figure 7.f: Results of 2023 of assessment of uncertainties for Charpy impact test, Lateral expansion *LE*.

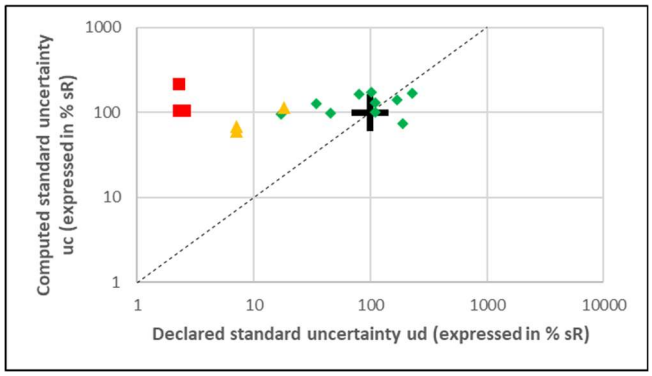


Figure 7.g: Results of 2023 of assessment of uncertainties for shear force of a welded knot.

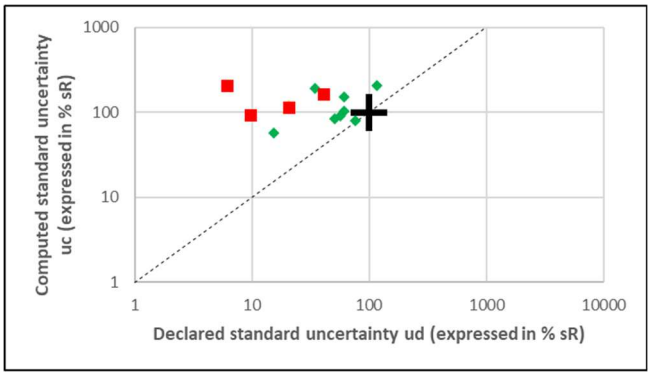


Figure 7.h: Results of 2023 of assessment of uncertainties for tensile test, Elongation at rupture A_{5d}.

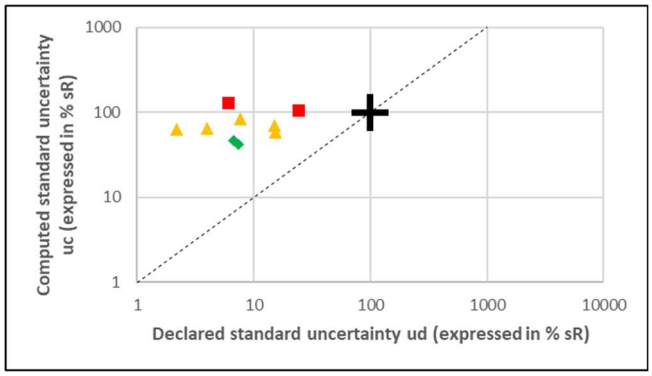


Figure 7.i: Results of 2023 of assessment of uncertainties for speed of intergranular corrosion in a Huey test on a stainless steel.

It can be observed from these results that:

- The statements of § 3.2 are confirmed from this extended series of results. The test methods for which CRM exist or are mainly metrological are those for which u_i are the best determined. Among those for which CRM exist (i.e. chemistry, hardness, tensile testing and Charpy), non-destructive ones are those for which u_i are the best determined;
- Standard deviations of s_ζ is a good parameter to determine how far the method is difficult with respect to determining correctly the associated uncertainties: the greatest s_ζ is, the lower the average u_i/s_R is, the more is the proportion of alerts, i.e. the more the participants u_i are underestimated;
- The more the determination of u_i is difficult, the less the proportion of laboratories determine them;
- A special notice needs to be made for the shear test on welded fabrics, which is obviously a full "technological" method, but for which a large proportion of laboratories determined u_i from results of ILC to which they participated formerly. We can then see that, in that case, using results of ILC is the only way to get correct determinations of uncertainties.

12 Conclusions

A survey of results of assessing uncertainties during ILC operated by CompaLab showed that most of uncertainties determined by participants are significantly underestimated. With respect to this, 3 categories of test methods can be distinguished:

- Test methods that are mainly metrological and/or for which CRM are available (typically chemistry);
- Test methods that include a significant part of qualitative sources of uncertainties (typically tensile tests);
- Test methods which uncertainties are mainly governed by qualitative sources (typically corrosion tests).

Unsurprisingly, uncertainties are globally well determined for some of the test methods of the 1st category while they are globally underestimated by a factor 1 to 10 or more for some of the test methods of the 3rd category.

This situation seems to come from a massive choice of laboratories to use GUM method B to determine their uncertainties, whatever the test method consists of. Several reasons may explain this situation:

- Until twenty years ago, most of laboratories were not familiar with techniques of determination of uncertainties. They were then inclined to subcontract the job to consultants, that mainly come from metrology where GUM method B is usually effective;
- GUM method B can be implemented within a single laboratory while other methods need direct or indirect (i.e. using RM) collaboration between several laboratories.

However, GUM methods, particularly the method B is effective in the field of metrology (where standards of calibration are always available) but is not when significant qualitative sources of uncertainty are present, what is quite common is lab testing. Because of that:

- The issue of bias of more importance in lab testing than in calibration;
- The laboratories are inclined to consider as negligeable some qualitative sources of uncertainty just because their contributions is hard to determine by this method.

Moreover, some issues more or less specific to testing are not addressed in GUM (bias of the test method, effect of material, inner homogeneity of test specimens, effect of range of measurands, preparation of test specimens, …) and GUM lacks guidance about these issues.

On the other hand, information coming from other sources, typically results coming from ILC and results from programs of surveillance of quality implemented by the laboratory can be re-used directly or as raw material for GUM method A experiments to determine uncertainties in a way that better takes into account the issues listed here upper, and therefore, provide quite better estimates of uncertainties. Moreover, re-using such data requests a significantly lower amount of internal resources (time and money) than implementing the GUM method B.

When it is of importance that uncertainties are well determined, large collaborative GUM method B (i.e. specifically designed ILC) should be organised, which results may also be used afterwards in a very effective programs of internal surveillance of quality of testing as requested for accreditation of laboratories.

Determining uncertainties should always begin by:

- A clarification about what is the intended use of uncertainties to be determined;
- A collection of publicly or internally available information concerning the precision of testing (in particular from former results of ILC);
- An estimation of the type of test method (highly qualitative or not)

The most appropriate method to determine uncertainties highly depends on the answers to these questions. And in most cases, the answer is not “GUM method B”!

13 References

[1] ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories

[2] BIPM, JCGM 100, Evaluation of measurement data - Guide to the expression of uncertainty in measurement, 2008, DOI: https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6?version=1.12&t=1696944486074&download=true
also published by ISO under the reference ISO/Guide 98.

[3] ISO 13528:2022, Statistical methods for use in proficiency testing by interlaboratory comparison

[4] BIPM, JCGM 200, International vocabulary of metrology – Basic and general concepts and associated terms, 2008, DOI: https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1?version=1.16&t=1659082802818&download=true
also published by ISO under the reference ISO/Guide 99.

[5] ISO 3534-2: Statistics – Vocabulary and symbols – Part 2: Applied statistics

[6] ISO 5725-1:2023: Accuracy (trueness and precision) of measurement methods results – Part1: General principles and definitions

[7] ISO 6507-1:2023: Metallic materials – Vickers hardness test – Part 1: Test method

[8] ISO 5725-2: Accuracy (trueness and precision) of measurement methods results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

[9] ASTM E691:22: Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

[10] ISO 33405:2017: Reference materials — Approaches for characterization and assessment of homogeneity and stability

[11] ISO 21748:2017: Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation

[12] ISO 5725-3:2023 Accuracy (trueness and precision) of measurement methods results – Part 3: Intermediate precision and alternative designs for collaborative studies

[13] ISO 3:1973: Preferred numbers – Series of preferred numbers

[14] ISO 33406:2024: Approaches for the production of reference materials with qualitative properties

[15] ISO 16269-6:2014 Statistical interpretation of data – Determination of statistical tolerance intervals

[16] ISO 2854:1976 Statistical interpretation of data — Techniques of estimation and tests relating to means and variances

[17] David Luengo, Luca Martino, Mónica Bugallo, Víctor Elvira and Simo Särkkä, "A survey of Monte Carlo methods for parameter estimation" EURASIP Journal on Advances in Signal Processing, Article 25, May 2020
DOI: https://doi.org/10.1186/s13634-020-00675-6

[18] BIPM, JCGM 101, Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method, 2008
DOI: https://www.bipm.org/documents/20126/2071204/JCGM_101_2008_E.pdf/325dcaad-c15a-407c-1105-8b7f322d651c?version=1.15&t=1696950361579&download=true

Annex:
Examples of implementing different methods for the determination of uncertainties

Examples of publicly available results of ILC

Chemistry of metals: ISO/TR 9769:2018 - Steel and iron - Review of available methods of analysis

Many ASTM standards related to test methods provide information about repeatability and reproducibility from ILC, including ASTM A90, ASTM B487, ASTM B499, ASTM C736, ASTM C792, ASTM C1135, ASTM D522, ASTM D523, ASTM D638, ASTM D696, ASTM D785, ASTM D790, ASTM D792, ASTM D882, ASTM D1005, ASTM D1505, ASTM D2202, ASTM D2240, ASTM D2457, ASTM D2584, ASTM D2794, ASTM D3359, ASTM D3418, ASTM D3895, ASTM D4366, ASTM D5630, ASTM D6980, ASTM E8, ASTM E10, ASTM E18, ASTM E21, ASTM E23, ASTM E45, ASTM E92, ASTM E96, ASTM E112, ASTM E228, ASTM E308, ASTM E384, ASTM E399, ASTM E606, ASTM E647, ASTM E698, ASTM E793, ASTM E831, ASTM E930, ASTM E1131, ASTM E1356, ASTM E1641, ASTM E2041, ASTM E2550, ASTM G1, ASTM G28, ASTM G48.

Attention needs to be paid on the fact that ASTM E177 (§ 10) defines r (repeatability) and R (reproducibility) as values that approximate 95% of pairs of test results obtained in respectively in repeatability or reproducibility conditions do not overcome. Consequently, $r = 2{,}8.s_r$ and $R = 2{,}8.s_R$. Basics for that is provided in the same reference.

Example of random design of experiments for a performance of full GUM method A

Example of a laboratory preparing itself the test specimens, using 2 testing methods, 8 testing machines and employing 5 operators. It was decided that:

- *2 types of material (average and difficult);*
- *5 levels of measurands;*

Should be included in the experiment. The design of experiments should then include:

- *2 testing methods;*
- *2 types of material (average and difficult);*
- *5 levels of measurands;*
- *3 sources of preparation of test specimens (its own facilities plus, as far as possible, 2 external sources);*
- *8 testing machines;*
- *5 operators;*
- *2 environmental conditions (i.e. performance of tests at 2 different periods of the year);*
- *2 repetitions, if repeatability SD is needed.*

That is to say a total number of 2x2x5x3x8x5x2x2 = 9600 tests. This amount of testing is technically and economically impossible to implement. Moreover, RM are not available for all combinations of types of material and measurands. It is then decided to use a random design of experiment that takes into account:

- *The actual proportions of combination of testing methods – testing machines – levels of measurands – operators that the laboratory uses for producing the test results;*
- *Use weighting coefficients to make the experiment more representative to the actual operations of the laboratory. For doing it, coefficients of 0,3 to 3 were attributed to each combination, with 1 for average combinations;*
- *Levels of measurands 1 – 3 – 10 – 30 – 100. All testing machines are not able to address all levels of measurands;*
- *Need to use RM, but only on 20% of test results (enough to get acceptable accuracy on the determination of bias). Moreover, RM are available only on "average material", measurand level 3 and 30;*

- *Need to determine the repeatability, i.e. need of repetitions, but only on 20% of combinations (enough to get acceptable accuracy of determination of s_r).*

The matrix of combinations and of weighting were defined as follows:

Source of uncertainty	Occurrence	Weight
Type of material	Average	0,9
	Difficult	0,1
Preparation	Internal	0,333
	External 1	0,333
	External 2	0,333
Environmental conditions	Winter	0,5
	Summer	0,5
Repetitions	With	0,2
	Without	0,8

Testing method	Testing machine	Operator	Measurand	Weight
A	1	Albert	1 to 30	0,5
A	1	Benoît	1 to 30	2
A	1	Camille	1 to 30	0,5
A	2	Albert	3 to 100	2
A	2	Benoît	3 to 100	0,5
A	2	Daniel	3 to 100	0,5
A	2	Etienne	3 to 100	0,5
A	3	Albert	1 to 1000	0,3
A	3	Benoît	1 to 1000	0,3
A	3	Camille	1 to 1000	0,3
A	3	Daniel	1 to 1000	0,3
A	3	Etienne	1 to 1000	0,3
A	4	Daniel	10 to 100	2
A	4	Etienne	10 to 100	0,5
B	4	Daniel	10 to 100	0,5
B	4	Etienne	10 to 100	2
A	5	Albert	3 to 30	0,3
A	5	Camille	3 to 30	0,3
B	5	Albert	3 to 30	1
B	5	Benoît	3 to 30	0,5
B	5	Camille	3 to 30	2
B	5	Daniel	3 to 30	1
B	5	Etienne	3 to 30	0,5
A	6	Alfred	3 to 100	0,5
A	6	Benoît	3 to 100	0,5
A	6	Camille	3 to 100	0,5
B	6	Alfred	3 to 100	0,5
B	6	Camille	3 to 100	3
B	7	Daniel	3 to 30	2
B	7	Etienne	3 to 30	1
B	8	Daniel	10 to 100	1
B	8	Etienne	10 to 100	3

This makes 118x2x3x2 = 1416 possibilities among them 100 will be randomly chosen, among which 25 will be randomly chosen to be repeated in order to compute s_r. The tests on "average materials" measurands 3 and 30 will be performed on RM in order to determine the bias.

The weighted random selection can be achieved with the following steps:

- *Build up the matrix of all possible combinations including the global weight for each of them (this weight is obtained with the multiplication of all individual weights, example: the combination "average – internal – winter – A – 1 – Albert – 1" has a global weight of w = 0,9x0,333x0,5x0,5 = 0,075);*

- *Build up the weighted matrix, which repeats each combination in a weighted random number of times. A multiplication factor* m *(equal for all combinations) may be necessary to get appropriate integer number for each combination (example with a random number 0,44,* w *= 0,075 and* m *= 100, 0,44x0,075x100 = 3,3 to be rounded to 3);*
- *Select randomly the desired number of combinations to test in this latter matrix.*

An example of resulting design of experiment of 100 combinations to test, among which the 25 first are to be repeated, is provided here after.

Testing method	*Testing machine*	*Operator*	*Measurand*	*Type of material*	*Preparation*	*Environmental conditions*
A	1	Albert	10	Difficult	External 2	Summer
A	2	Albert	10	Average	External 1	Winter
A	1	Camille	10	Average	External 1	Winter
A	3	Benoît	10	Average	External 2	Summer
B	6	Camille	10	Average	External 1	Winter
B	8	Etienne	30	Average	External 1	Summer
A	3	Etienne	100	Average	External 1	Summer
B	8	Etienne	10	Average	Internal	Summer
A	3	Camille	10	Average	External 2	Summer
B	8	Daniel	100	Difficult	External 2	Winter
B	4	Daniel	100	Difficult	Internal	Summer
A	5	Albert	10	Difficult	External 2	Winter
B	8	Daniel	10	Average	External 1	Winter
B	8	Etienne	10	Average	Internal	Summer
A	1	Albert	30	Average	Internal	Summer
A	4	Etienne	100	Average	External 1	Winter
A	1	Albert	10	Difficult	External 2	Winter
B	5	Camille	3	Average	External 1	Winter
B	6	Camille	100	Average	External 2	Winter
B	7	Etienne	10	Average	External 2	Summer
B	7	Daniel	10	Average	External 2	Winter
B	6	Camille	3	Difficult	External 2	Summer
B	5	Camille	10	Average	External 1	Summer
B	5	Camille	3	Average	Internal	Summer
B	8	Daniel	30	Average	External 1	Summer
A	2	Benoît	100	Average	External 2	Summer
B	6	Alfred	3	Average	External 2	Winter
A	1	Benoît	1	Average	External 2	Winter
B	7	Etienne	3	Average	External 1	Summer
B	8	Daniel	10	Average	External 2	Summer

Testing method	*Testing machine*	*Operator*	*Measurand*	*Type of material*	*Preparation*	*Environmental conditions*
B	8	Daniel	100	Average	Internal	Summer
B	8	Etienne	30	Average	External 1	Summer
A	2	Albert	30	Difficult	Internal	Winter
B	6	Camille	30	Average	External 1	Summer
B	8	Etienne	10	Average	External 2	Summer
A	3	Albert	10	Average	External 1	Summer
B	6	Camille	3	Difficult	External 2	Summer
B	5	Camille	10	Average	Internal	Winter
A	2	Daniel	30	Average	External 1	Summer
B	7	Etienne	30	Average	External 1	Summer
B	6	Camille	100	Difficult	External 2	Summer
B	5	Benoît	30	Average	External 2	Winter
B	7	Etienne	30	Average	Internal	Winter
A	4	Etienne	30	Difficult	External 1	Summer
B	5	Camille	10	Average	External 1	Summer
A	3	Etienne	100	Average	External 2	Winter
B	6	Camille	10	Average	Internal	Summer
A	5	Camille	30	Average	External 1	Summer
A	3	Etienne	100	Average	External 2	Winter
A	5	Camille	30	Average	External 2	Winter
A	2	Albert	3	Average	External 1	Winter
A	1	Benoît	3	Average	External 2	Winter
A	4	Etienne	100	Average	External 1	Winter
A	6	Camille	30	Average	External 2	Winter
B	8	Daniel	30	Average	External 1	Summer
A	1	Benoît	1	Average	External 2	Winter
A	5	Camille	30	Average	Internal	Winter
B	7	Etienne	3	Average	Internal	Summer
B	4	Daniel	10	Average	Internal	Winter
A	3	Camille	30	Difficult	Internal	Winter

Testing method	*Testing machine*	*Operator*	*Measurand*	*Type of material*	*Preparation*	*Environmental conditions*
B	6	Camille	3	Average	External 1	Winter
B	7	Daniel	3	Average	Internal	Winter
B	6	Alfred	100	Average	External 2	Summer
A	5	Albert	10	Difficult	External 1	Winter
B	8	Daniel	10	Average	External 1	Summer
B	6	Camille	100	Average	External 1	Summer
B	7	Etienne	10	Average	Internal	Summer
A	2	Daniel	30	Average	External 2	Summer
B	5	Albert	30	Average	Internal	Winter
B	5	Benoît	30	Average	Internal	Summer
A	1	Albert	30	Average	Internal	Summer
B	8	Etienne	30	Average	External 1	Summer
B	4	Etienne	30	Average	External 2	Winter
A	3	Etienne	100	Average	External 1	Summer
B	4	Etienne	10	Average	External 2	Winter
A	6	Benoît	30	Difficult	External 2	Winter
B	4	Daniel	10	Average	External 1	Winter
B	8	Etienne	30	Average	Internal	Summer
A	1	Albert	30	Difficult	Internal	Winter
A	2	Albert	30	Average	External 1	Winter

Testing method	*Testing machine*	*Operator*	*Measurand*	*Type of material*	*Preparation*	*Environmental conditions*
A	6	Alfred	100	Average	External 2	Summer
B	4	Etienne	10	Average	External 2	Summer
A	2	Albert	30	Average	Internal	Summer
B	8	Daniel	100	Difficult	External 1	Winter
B	8	Daniel	100	Average	External 2	Winter
B	5	Camille	10	Average	Internal	Winter
A	1	Benoît	1	Average	External 1	Winter
A	6	Alfred	100	Difficult	External 2	Winter
A	5	Albert	3	Average	External 2	Winter
A	6	Alfred	3	Average	Internal	Summer
A	3	Benoît	3	Average	External 1	Summer
B	7	Daniel	10	Average	External 2	Winter
A	3	Benoît	100	Average	External 1	Summer
A	2	Albert	30	Average	External 2	Winter
B	5	Daniel	3	Average	External 2	Winter
A	2	Benoît	10	Average	Internal	Winter
B	5	Etienne	3	Average	Internal	Winter
A	1	Camille	3	Average	External 1	Winter
B	6	Alfred	100	Average	Internal	Summer
A	1	Albert	3	Average	External 2	Summer

The resulting total amount of combinations per source of uncertainty is as follows:

Source of uncertainty	***Occurrence***	***Total number***	***Total number of repeated ones***
Type of material	Average	84	19
	Difficult	16	6
Preparation	Internal	27	5
	External 1	35	10
	External 2	38	10
Environmental conditions	Winter	50	11
	Summer	50	14
Method	A	46	10
	B	54	15

Source of uncertainty	*Occurrence*	*Total number*	*Total number of repeated ones*
Machine	1	12	4
	2	10	1
	3	10	3
	4	9	2
	5	17	4
	6	17	3
	7	9	2
	8	16	6
Operator	Albert	17	5
	Benoît	12	1
	Camille	23	8
	Daniel	18	5
	Etienne	24	6
Measurand	1	3	0
	3	18	3
	10	30	14
	30	29	3
	100	20	5

Example of study in 2 steps

An alternate way for the upper example would be to implement two steps, as follows:

1. *Perform 8 to 10 tests for each source of uncertainty and compute the related bias and random error;*
2. *Select the one or two or three main sources of uncertainty and perform a series of tests on random combinations limited to the sources of uncertainty that mainly contribute to the global uncertainty.*

The design of experiment could typically be as follows:

1. *8 tests on average material and 2 on difficult material;*
2. *3 tests on each of the sources of preparation of test specimens;*
3. *5 tests for two testing moments;*
4. *5 tests for each test method;*
5. *1 test for each testing machine;*
6. *2 tests for each operator:*
7. *2 tests for measurand "3", 3 tests for measurand "10", 3 tests for measurand "30", 2 tests for measurand "3", on RM for measurands 3 and 30.*

For each of them, all other sources of uncertainty shall be same occurrence, as far as possible (for test method, none of the machines and operators may be capable to produce test results for both methods).

Examples of implementations of GUM Method B

Following documents enclose examples of determinations of uncertainties using the GUM Method B:

- *GUM [2] Annex H: several examples among which one related to a test method (Rockwell hardness C);*
- *ISO 148-1, concerning the Charpy impact test;*
- *ISO 4545, concerning the Knoop hardness test;*
- *ISO 6506-1, concerning the Brinell hardness test;*
- *ISO 6507-1, concerning the Vickers hardness test;*
- *ISO 6508-1, concerning the Rockwell hardness test;*
- *ISO 6892-1, concerning the tensile test of metals;*
- *ISO/TR 15263, concerning the tensile test of metals.*