UNCERTAINTY IN MEASUREMENT SCIENCE

L.Mari [(*)], *G.Zingales* [(**)]
[(*)] Libero Istituto Universitario C.Cattaneo,
C.so Matteotti 22, 21053 Castellanza (Va), ITALY
[(**)] Dipartimento Ingegneria Elettrica - Università di Padova
via Gradenigo 6/a, 35131 Padova, ITALY

Abstract: The spread of the measurement results, taken in supposedly identical conditions, is a well known fact occurring in all the experiments. The traditional interpretation is based on the assumption of a "true value" for the measurand and of "errors" that cause the observed variability. More recently the non-exactness of the measurement results has been understood in terms of their uncertainty, and a specific procedure to derive the uncertainty of measurands in the case of indirect measurement has been proposed by several standardization bodies, as formalized in the "Guide to the expression of uncertainty in measurement". This work is aimed at critically analyzing the foundations of the notion of uncertainty in measurement and its novelty with respect to the previous conceptions, also highlighting the reasons of both adequacy and insufficiency of the mentioned Guide.

Keywords: uncertainty, foundations of measurement

1. NATURE AND REPRESENTATION OF THE UNCERTAINTY

1.1. The uncertainty in measurement is a relatively recent issue …

Although men measure since long time and with many different purposes, the need to evaluate and formally express the non-exactness of the measurement results is a relatively recent issue. At least two distinct reasons can be suggested for this apparent anomaly. On one hand, whenever it is agreed to report the results with a sufficiently low number of significant figures, as it is customary in commerce and many practical purposes, the measurement appears as actually repeatable. On the other hand, the traditional metaphysical belief that "the numbers are in the world", strengthened by the increasing success of the infinitesimal analysis during the last three centuries, grew in the myth of the single-valuedness of the measured quantities. From these complementary reasons the apparently counterintuitive conclusion can be reached that the experimental variability of the measurement results is not a generally observed (or at least reported …) fact.

It is worth to remark that the topic of the evaluation and representation of the uncertainty of measurement results has recently received a specific attention by a wide audience, due to the publication in 1993 of the "Guide to the expression of uncertainty in measurement" (GUM) [1], the result of the coordinated efforts of several international standardization bodies (cf. Section 2.2). While aimed at generally analyzing the topic, our analysis will then maintain the GUM as a reference, to discuss its usefulness and to highlight its merits and critical points.

1.2. The reasons

The fact that the repeatability of the measurement depends on the number of significant figures reported in its results highlights the relevance of the linguistic component of the operation. Measurement results are indeed specific linguistic descriptions and as such they inherit the general properties of these descriptions, in particular the existence of a relation between the degrees of specificity and certainty of the description itself.

Consider, for example, the two descriptions A="this is a book" and B=" this is a 120 page book", both pronounced while indicating the same object. B logically entails A and therefore

is more specific than it. As a whole, B is however less certain, i.e., more uncertain, than A. The less the descriptions produced by an observation are specific, the more the observation itself is repeatable. The need to recognize and formalize the presence of a non null uncertainty arises only when the descriptions reach a sufficiently high degree of specificity.

In this view the history of the past centuries exhibits a seemingly paradoxical situation: «when we look to earlier times [before XIX century] for the connection between numbers stated and observations made, historians are even more puzzled. Galileo may have been the first to think about averages, and it was a long time until the arithmetic mean -- averaging -- was a commonplace for experimenters. Gauss had provided a theory of error by 1807, and astronomers made use of that. Although all modern physical measurements demand an indication of error, physics outside of astronomy did not report estimates of error until the 1890s (or later)» [2].

Measurement results have been thus traditionally reported as both highly specific (i.e., as numbers with many significant figures) and completely certain, and this although in a typical continuous probability distribution any single value would have only an infinitesimal degree of certainty.

1.3. The hypothesis of the true value

Among the different hypotheses that can be adduced to justify a posteriori this situation, one can be highlighted as particularly meaningful: from the methodological intent of making the scientific results quantitative, an idealistic philosophy has believed to derive the ontological assumption that measurable quantities are characterized by a single, thus maximally specific, “true value”. As a consequence, any discrepancy between the measurement results and the measurand true value has been considered as an “error”, accounting for this difference and as such minimized as far as possible. It is now more and more commonly recognized that these concepts cannot be maintained against an operational analysis requesting to identify a procedure aimed at determining the true value of a given measurand [3]. The more and more accepted shift from “error” to “uncertainty” is then more than purely terminological, and highlights the recognition of a different epistemology for measurement. In this regard the position of the GUM seems to be agnostic («the terms “value of a measurand” and “true value of a measurand” are viewed as equivalent» so that «the term “true value” is not used»), but is substantially still classic (any measurand is «characterized by a unique, invariant value»). The criticality of this position emerges when considering the following two statements:

“at the instant of the measurement the thing is in a definite state”

and:

“at the instant of the measurement the measurand has a definite value”

Traditionally such statements would be plausibly considered as synonymous, whereas their conceptual distinction is a fundamental fact of metrology: the former represents a basic assumption of the measurement (neglecting here the issues related to the role of the measurement in quantum mechanics), while the latter is epistemically unsustainable and however operationally irrelevant. Measurement results are linguistic, and not empirical, entities: what in the measurement is determined, and therefore considered pre-existing, is the thing state, and not the measurand value, that is instead assigned on the basis of the instrument reading and the calibration information.

1.4. Intrinsic uncertainty and single-valuedness

The fact that measurement results are linguistic constructs has a second, fundamental consequence. Any measurement result includes the specification of a measurand, that must be

thus defined a priori and then suitably expressed. This linguistic specification is only a partial description of the thing state, resulting in an under-identification of such a state, that maintains the unavoidable vagueness accounting for the essential distinction of «the realm of things we say from the realm of things we do» [4] (operatively the reasons are various, including the lack of specification and/or control of all the relevant influence quantities and an insufficiently documented and/or controlled traceability chain). It is precisely the presence of such a vagueness, usually called "intrinsic uncertainty" in metrology, that excludes any meaningfulness for the assumption that the measurand «to be invariant and to have a unique, fixed value» as instead the GUM assumes.

On the other hand, this ontological conclusion has no direct implications with the linguistic remark that a value for the measurand be possibly formalized as a singleton (i.e., a subset whose cardinality is 1), an option whose usefulness is manifest whenever some symbolic processing is needed (the mathematics of intervals, or subsets, or fuzzy subsets, … is overwhelmingly more complex than the usual one for scalars). Again, this is related to linguistics, and not to empirical facts: part of the measurand definition, to be made before the measurement, is indeed the decision of which set of values to adopt. The complete specificity of the measurement result, i.e., the fact that the assigned value is a singleton, depends on such a set (so that if the set {"high", "medium", "low"} is chosen, a measurement result specifying a "medium" length would be maximally specific …).

1.5. The expression of non-exact measurement results

Once acquired, the empirical information on the measurand can be represented by suitably modulating its specificity and certainty, whose combination formalizes the non-exactness degree of the measurement result. Therefore the same expression, as the traditional $x \pm y$, admits two distinct, and actually opposite, meanings:

M1: the measurand value, the singleton x, is maximally specific but uncertain, with uncertainty y;

M2: the measurand value, the interval $[x-y, x+y]$, is not completely specific but certain

(an observation on M2, that is sometimes misunderstood in its interpretation. According to M2 the measurement result is the interval as a whole, and as such its certainty is considered so high, e.g., 95% or higher, that it is dealt with as practically certain. Given this interval, each of the singletons belonging to it is actually uncertain, and maximally specific. But this implies a swap to M1).

The GUM allows in principle both these interpretations. While focusing on M1 and in this context presenting its basic results (cf. Section 2.3), it introduces a simple procedure to re-express a measurement result in the form of a confidence interval, i.e., according to M2 (cf. Section 2.4). The qualifying (and weak, in some sense) point of this approach is the complete subordination of the M2 representation to the M1 one: indeed the GUM assumes that an interval can be computed only from a couple (singleton, uncertainty value). This prevents the adoption of further, non statistical, representations, such as those based on fuzzy set theory.

The fact that the measurement results are expressed as non-exact (either specific but uncertain, or non-specific but certain, or both non-specific and uncertain) has a fundamental consequence on the way they can be compared with each other to ascertain whether two results identify the same measurand or not. While in the exact case such a comparison is performed in terms of the identity of the measurement results, in the non-exact case this criterion appears too exigent. The compatibility of two results is indeed guaranteed by weaker conditions (cf. Section 3.1), that maintain the reflexivity and symmetry but in general relax the transitivity of the comparison relation [5]. This highlights an inadequacy of the current

axiomatic foundations [6], that formalize measurement as a homomorphism between relational structures, i.e., a function mapping empirical states into symbols and preserving the empirical relations so that they are expressed by measurement results. Although the GUM does not deal with this issue, its practical importance, e.g., in standard comparison, is manifest.

A further insufficiency of the GUM and other approaches aimed at expressing the uncertainty in measurement (cf. Section 2.5) has to be remarked, related to their inability to cover the results of measurements performed in non-ratio (or interval) scales, as they appear, e.g., in ordinal evaluations. In such cases the very concepts of average, standard deviation, interval, … are not defined, and weaker, i.e., more general, procedures must be investigated.

1.6. The need for an ethic of metrology

It is not the empirical reality that imposes a value for the measurand, pre-existing to the measurement and determined by means of it (and then "true" as far as close to the "real" value). A measurand value must be assigned by the operator, by means of the suitable application of a measuring system, on the basis of the goals for which the measurement is performed, and it results adequate (and not "true") if it meets such goals, as expressed in terms of a given minimum specificity and maximum uncertainty. By means of the combination of these two parameters the operator is able to express an index of the quality of the measurement result. That is why no method for evaluating the measurement uncertainty can be a «substitute for critical thinking, intellectual honesty, and professional skill.» Indeed «the quality and utility of the uncertainty quoted for the result of a measurement ultimately depends on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value» [1]. The same, of course, holds more generally for the assignment of a measurement result, thus including the measurand value.

2. THE EVALUATION OF UNCERTAINTY

2.1. The theory of random errors

In the classical theory of errors any instrument reading x was considered as the result of a combination between the value x_0, that would have been obtained by a "perfect instrument", and the effect ε, a "random error" produced by several small different sources, due to the actual instrument performance, randomly varying both in sign and amplitude. When repeating the same measuring procedure, even with the same instrument and the same environment and operating conditions, different readings would be obtained. Because of this random behavior, a statistic model was assumed for ε, assigning to it a Gaussian probability distribution (PD), with zero mean. Then the mean $\mathrm{x} = \mathrm{E}(x_i)$, obtained by a finite number n of repeated readings x_i of the same measurand, was considered as the best attainable result to express the measurand value. Its dispersion was expressed by different parameters, ranging from the average of the absolute deviations $|x_i - \mathrm{x}|$ up to the estimated standard deviation (SD) of the mean $s = s(\mathrm{x})$, given by:

$$s = \qquad (1)$$

These formulas were not always used in practical applications, mainly because of a lack of sound foundation and general agreement on their usage. For example, some old textbooks noted that generally the unknown systematic errors were larger than the random ones and it was therefore «rarely useful to perform complicated calculations» (especially when made by hand or by a sliding rule) to evaluate random errors. On the contrary, in many cases the assumed Gaussian PD was used to characterize the random error ε as a function of s, corresponding to the amplitude of the interval around the mean value inside which the value

x_0 could be found, with a given probability (or "confidence level"). The most commonly used values have been a "probable error" of 0,68 s for a confidence level of 50%, an interval s for a confidence level of 68%, or 2 s for a confidence level of about 95%. In the general case a functional relation $y = f(x_1, x_2, \ldots, x_N)$ is known and the random errors $\varepsilon(x_i)$ are available, the random error $\varepsilon(y)$ of y was computed from the $\varepsilon(x_i)$ by considering $\varepsilon(y)$ as a differential of the function f, and the following expression was generally derived:

$$\varepsilon(y) = \qquad (2)$$

such a value being defined as the limit absolute error, describing the "worst possible case" in which all the sources were supposed to produce errors with the same sign.

It was however difficult to separate the sources of random errors from those of systematic ones, as the randomness of a source depends on the operating procedure: this is particularly important in the case of repeated readings of the same instrument. When successive readings are taken in a short time, so that the influence quantities can be supposed to remain constant, only a part of the instrument random errors are present (like resolution, quantization, noise, reading uncertainty, friction between moving parts, …). The concerned dispersion expresses in this case the repeatability of the instrument readings. When using an instrument under different conditions of its influence quantities, their effects had to be considered as further sources of random errors, superimposed to the previously mentioned ones. When comparing two different instruments of the same type, the possible differences in their calibrations should also be considered; in this case it was common to analyze the topic in terms of the reproducibility of the measurement results.

2.2. Establishment of a Standard

The necessity of a common way for expressing uncertainty was urgently posed when a strict co-operation was being established among the national calibration laboratories in Europe, about 25 years ago. To build up and maintain a mutual confidence between accreditation bodies and an equivalence for their calibration certificates required to have their uncertainties evaluated and expressed in a harmonized way. The work, initiated by CIPM, was later performed by a joint committee of experts from different international organizations involved in standardization (ISO, IEC, OIML, …). Its final result is the already mentioned GUM, first published in 1993 and later introduced as a Standard by each of the above mentioned organizations. While originally intended for calibration laboratories, the GUM is presently to be considered as the basis for expressing the results of any measurement performed in accordance with an international Standard.

As any document produced by a committee of experts, the GUM reaches useful practical solutions through compromises between different views. One of these pragmatic compromises may be found when comparing its specific procedure for evaluating uncertainty with the definition itself of uncertainty as «a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand», a definition that better corresponds to the classical region of reasonable doubt or empirical safety, represented by the "worst possible case" having less sharp borders, that has been used for a long time by scientists and technicians.

2.3. The GUM procedure

The starting point of the GUM is to consider all the physical quantities involved in a measurement (such as measurands, influence quantities, corrections, reference materials properties, manufacturer or reference data, …) as random variables, whose PDs are characterized by their SDs and to use such SDs as a basis for representing the uncertainty for the values of the PDs. Each measurement result is then expressed as a couple $<x,s(x)>$ where

the first term is a singleton assumed as the measurand value and the second term is its SD (the GUM calls it a "standard uncertainty"; for the sake of simplicity the common statistical term of standard deviation will be however maintained in the following), a representation based on what has been previously called the M1 interpretation (complete specificity and uncertainty).

In contrast to this formal homogeneity, a pluralistic view is assumed on the interpretation of the probability, and therefore on the procedure to obtain the SDs. The GUM recognizes indeed that the uncertainty of a measurement result can derive from two distinct areas: some uncertainties, designated "of type A", are computed by the statistical treatment of experimental data, obtained as the observed frequencies in repeated experiments; some other uncertainties, designated "of type B", are instead estimated by scientific judgements based on personal experience and a priori information and therefore describe a degree of belief on the occurrence of events. Operatively, from this distinction a suggestion is derived on the hypothesis to be preferred for the PD: Gaussian for type A and uniform for type B.

Although all the details about the GUM procedure should be left to a careful reading of its full text, some basic principles may be here recalled.

The core of the GUM procedure is the law of propagation of the variances (that the GUM interprets thus as a law of propagation of uncertainties) to be applied for derived measurement. Let Y be a quantity depending on N input quantities X_i, i=1,…,N, through the functional relationship:

$$Y = f(X_1,\ldots,X_N) \tag{3}$$

(we are following the usual convention according to which the quantities are denoted in upper case and the values they assume in lower case). In the case the measurands X_i are considered only uncertainly known, are therefore their values are expressed as $<x_i,s(x_i)>$, the issue arises of how to obtain a corresponding value $<y,s(y)>$ for Y. Consistently with its basic hypotheses, the GUM assumes that the measurand value y and its SD can be obtained by means of separate computations. While an estimate y of Y may be obtained by introducing the estimates x_i of the input quantities X_i in the model function f, in the simplest case of non-correlated quantities the variance $s^2(y)$ of y can be obtained from the variances $s^2(x_i)$ of its N input quantities x_i by the formula:

$$s^2(y) = c_i^2\, s^2(x_i) \tag{4}$$

The sensitivity coefficients c_i describe the extent to which the output estimate y is influenced by variations of the input quantities x_i. In most cases they can be evaluated from the first partial derivative of the model function f:

$$c_i = \tag{5}$$

Numerical methods or experiments may be used when the model function is not sufficiently known or it is strongly non-linear.

When two input quantities x_i and x_j are known to be to some extent correlated, the variance $s^2(y)$ shall consider also their covariance $C(x_i, x_j)$:

$$s^2(y) = c_i^2\, s^2(x_i) + 2\, c_i\, c_j\, C(x_i, x_j) \tag{6}$$

While all the terms of the first summation are positive, in the second summation the covariance may be either positive or negative. It may also be expressed by using the SDs $s(x_i)$ and $s(x_j)$ and the corresponding correlation coefficient r_{ij} ($| r_{ij} | \leq 1$):

$$C(x_i, x_j) = r_{ij}\, s(x_i)\, s(x_j) \tag{7}$$

A correlation should be generally evaluated through the operator knowledge and experience: it may occur because of the use of the same reference data, reference standard or measuring instrument, in determining both x_i and x_j.

The GUM gives only a suggestion about the processes to be considered as uncertainty sources; besides the sources already mentioned in the Section 2.1, they include:

- an incomplete definition of the measurand;
- an imperfect realization of this definition;
- the representativeness of the sample in respect of the defined measurand;
- the assignment of values to measurement standards and reference materials;
- the approximations and assumptions included in the measurement method and procedure;
- the assignment of values to constants and other parameters obtained from external sources and used in the data reduction algorithm.

It is therefore necessary to fully investigate the measurement process, in order to recognize the different possible uncertainty sources and to quantify their contribution. The result is a sort of uncertainty budget, from which the resulting uncertainty can be derived.

An important practical result of the GUM procedure is the standardization of the numerical expression of an uncertainty to at most two significant figures. Furthermore, the usual rules should be used for rounding, with preference to round up. EAL (European co-operation for Accreditation of Laboratories) [7] recommends to use the rounded up value if the opposite would reduce the value of uncertainty by more than 5%. Correspondingly the numerical value of the measurement result, in its final statement, should be rounded to the least significant figure involved in its assigned uncertainty.

The importance of these rules stands in the underlying consideration that an uncertainty value is only an estimate of a complex and partly unknown process and in imposing a corresponding limit to the significant figures also for the result.

2.4. From uncertainty to non-specificity

The GUM suggests the SD $s(y)$ resulting from the formulas (4) or (6) as the main parameter for evaluating the uncertainty of a measurement result. However, «to meet the needs of some industrial and commercial applications, as well as requirements in the areas of health and safety», an expanded (or overall) uncertainty $u(y)$ is defined, obtained by multiplying $s(y)$ by a coverage factor k:

$$u(y) = k\, s(y) \tag{8}$$

As previously considered, the adoption of a coverage factor implies more than a trivial multiplication, and corresponds to swap from one way to interpret the non-exactness of measurement, completely specific but uncertain results, to the opposite one, non-specific but certain results. Indeed, the meaning and the way to evaluate the factor k are clearly understood when a definite PD is attributed to the measurand y. The quantity $k\, s(y)$ may then represent the interval including a given percentage of this PD (e.g., as stated before, 95% of a Gaussian PD with SD s is included in the interval of $\pm 1{,}97\, s$ around its mean).

This concept may be extended to other forms of PDs, resulting from the combination of the PDs of different uncertainty components, when their degrees of freedom are known. In some cases to compute a value of k corresponding to given percentage of the PD can be much more difficult, even if appeal can be made to the Central Limit Theorem, to some hypotheses on the existing degrees of freedom and to some statistical rules for combining them. In most cases the resulting value for k, for a confidence level around 95%, is between 2 and 3: hence many specifications for laboratory practice recommend to express overall uncertainty using a standard value of k=2, or to specify the reason why other values have been adopted.

Nonetheless, the main point is not the procedure for computing k but the meaning that should attributed to it. While a SD s may be accepted as a quantitative information about uncertainty, the accurate definition of both a final PD and its confidence intervals seems to many people practically unfeasible, and however an inadequate way to express the non-

exactness of the measurement results, mainly because of the difficulties that arise in the formal treatment of non-specific entities such as intervals, for which no special importance could be given to the central value of the interval with respect to the other ones.

2.5. Other procedures for evaluating uncertainties

As every result obtained from compromises, the GUM procedure may be criticized from different, and sometimes opposite, viewpoints. On the one hand, it has been noted that several hypotheses on which such a procedure is more or less explicitly based are not always applicable. Examples are the hypotheses of a zero central value and of a symmetrical PD for all the corrections; the requirement of assigning a priori a PD to formalize the type B uncertainties, … On the other hand, the definition and evaluation of the confidence limits may appear too detailed, when the generally low level of a priori knowledge is taken into account. In evaluating uncertainty sources of type A, indeed, the values of their SDs resulting from a limited number of experimental data should be considered only as rough estimates, thus with large uncertainties associated with them.

With the goal of overcoming these difficulties and encompassing in a unique conceptual and formal framework the treatment of non-exactness in measurement, several authors [8,9] have presented a different uncertainty theory, based on Bayesian statistics. While the GUM recognizes a parallel status for objective and subjective criteria in the uncertainty evaluation by making both type A and type B uncertainties coexistent, this competing approach emphasizes the importance of the operator's knowledge on the measured system for such an evaluation. When setting up the devices to perform a measurement, some a priori knowledge on the measurands (like their possible ranges of value, frequency spectra, …) must indeed be available to select the scale, sensitivity, bandwidth, … of these devices. The dependence of the quantities to be measured from external influence quantities and the presence of noise and interference should also be evaluated a priori, at least in statistical terms.

A random variable, named "estimator", is then associated with every input quantity playing a role in the measurement, regardless of whether or not this quantity generates random data. Any estimator is associated with its uncertainty, expressed, as previously indicated, by the SD of the PD attributed to it, that represents the incomplete information available.

The estimators replace the input quantities in the mathematical model of the measurand and shall therefore satisfy the model relationships. From these hypotheses, the distribution can be found that maximizes the likelihood of the measurand for a given result (principle of maximum information entropy), and its expected value and uncertainty can be computed.

In the simplest cases the results of such a procedure are not so different from those obtained according to the GUM prescriptions, but this approach can be successfully used also in special complex cases as non-linear, multi-measurand and function measurements.

2.6. Comments to the above procedures

The various procedures described in the above Paragraphs may be considered as mathematical models for representing the measurement process.

Although the currently available computing resources allow to deal with models much more complex than the ones effectively manageable until the recent past, we believe that the simple procedure specified in the GUM is at present adequate for industrial measurements. Research efforts should be devoted rather to clarify the application of such a procedure to practical cases, helping to determine the parameters to be introduced in its formulas, to avoid confusion and too pessimistic or too optimistic conclusions. For instance, to estimate a value for the uncertainty due to a measuring instrument would require a calibration performed under the combined ranges of the influence quantities. For many instruments, however, it is a common practice to have standards stating "error limits" separately for the reference

conditions of the influence quantities ("intrinsic error") and for their variations in the full range of a single influence quantity. This makes the acceptance tests simpler on these instruments, but makes very difficult to assess a reasonable value of the overall uncertainty in practical operations: indeed, the effects of one or more influence quantities could be much lower than these limits, and the possible correlation between the effects of different influence quantities is generally unknown. The GUM procedure may be of easier application for instrument calibration, when the standard assumed as reference is well known by the calibration laboratory and the terms of the uncertainty budget may be stated according to the operator capacity and authority.

Industrial tests are much more difficult to deal with. The uncertainties due to the test conditions are not always known in these cases, although they can be of great importance for the final result. Furthermore, several different values are usually obtained by readings on the same instrument, so that a large correlation factor may exist between the results. For instance, in computing the ratio or difference between two different but successive readings of the same instrument, only the sources affecting its repeatability and the uncertainty due to its non-linearity should be considered, all the effects of the other sources being almost cancelled by their strict correlation. It is not possible to distinguish these two types of uncertainty sources with the common manufacturer instructions.

It is difficult to obtain a satisfactory uncertainty budget by a debate between manufacturer and customer, especially when large economic interests are involved. This situation will hopefully change in a next future, when instrument manufacturers will fully adapt their specifications to the GUM requirements and the international Standards will provide clear statements for every test procedure, to have unquestionable, even if conventional, rules for the values to be introduced in an uncertainty budget, so to avoid discrepancies in the evaluation of a result.

It should in any case be remembered that an uncertainty value cannot be strictly determined but it is always estimated. As the assumption of the measurand "true value" has been rejected, it is obvious that a "true value" for the uncertainty attributed to the measurand value cannot be defined.

3. CONSEQUENCES

3.1. Compatibility

Besides its nature of quality index for measurement, uncertainty has important applications whenever a decision has to be taken related to the results of non-exact measurements. The most important practical cases are the statements of compatibility (or conformity) and of tolerance. These issues have not yet been fully covered by international Standards, so that the two terms are not even present in the "International vocabulary of basic and general terms in metrology" [10].

In general, two measurement results are said to be compatible when they can be attributed to the same measurand; the simplest condition for stating their compatibility is obtained by considering them as intervals, i.e. with a M2 interpretation, and requiring their non-null intersection. For two measurement results:

$$x_1 = m_1 \pm u_1 \qquad (9)$$
$$x_2 = m_2 \pm u_2$$

the compatibility condition is therefore:

$$|m_1 - m_2| \leq u_1 + u_2 \qquad (10)$$

(it should be noted that, due to its generality, this condition can be easily extended to ordinal, and even nominal, cases).

As a reduction in uncertainties may cancel the presence of common values, any compatibility statement is an ascertainment of the fact that the available data do not allow to distinguish the results. Compatibility statements should therefore be associated with the uncertainty levels by which they have been derived.

Compatibility is relevant when different measuring procedures or different measuring systems are employed on the same measurand. This is the case of the instrument checks performed by calibration laboratories and of measurements performed with the same procedure and similar instruments on two different test objects to demonstrate the equivalence of their parameters. The differences among these two applications will be shown in the following.

It is important to remark that, in principle, compatibility is not transitive, i.e., from the compatibility between x_1 and x_2 and between x_2 and x_3 the compatibility between x_1 and x_3 cannot be in general assumed.

3.2. Rules for compatibility

Besides the simple formula (9) the compatibility condition between two measurement results may also be expressed using statistical criteria [11,12] by a relation like:

$$| m_1 - m_2 | \leq f(s_1, s_2) \tag{11}$$

where s_1 and s_2 are the SDs of the PDs assumed for the values of M and from which the extended uncertainties u_1 and u_2 have been obtained. The previous formula may be reduced to:

$$| m_1 - m_2 | \leq \beta\, s_d \tag{12}$$

when an appropriate value s_d is defined as the SD for the difference $d = m_1 - m_2$. For non-correlated measurements:

$$s_d^2 = s_1^2 + s_2^2 \tag{13}$$

The form of function f and the value of coefficient β depend on the procedure followed in defining the compatibility criterion.

3.3. The effect of traceability

When evaluating the compatibility between two results, a special consideration should be given to the traceability of the concerned instruments, i.e., to the fact that both their outputs should derive, through an uninterrupted chain of calibrations, from the international standard of the measured quantity [13].

In a simplified view, a calibration consists in creating a correspondence between the value x_0 attributed to a standard, and the value m_0 given by the instrument, in a well defined set of its operating conditions. To attain this correspondence may require to operate regulations, to use scale factors, to introduce corrections, …

A second step to make an instrument suitable for accredited measurements is to assign to it an appropriate (extended) uncertainty u such that whenever, in any operating condition included in its rated range of use, it is compared again to the standard x_0, this value should be included in the measurement result. This is in fact the condition for passing a verification: otherwise the instrument should be re-calibrated. Such a process is however not intended for determining the value of the measurand, that is supposed to be already known, but only for checking whether the instrument has maintained its metrological characteristics.

Two instruments, complying with the above mentioned conditions of calibration, uncertainty and stability, will then always give compatible results when compared to the

standard x_0, or when applied to the same measurand having, like a standard, a negligible intrinsic uncertainty. As this principle can be applied to any pair of instruments, it may be concluded that, under these conditions, compatibility is transitive.

To have two compatible measurement results does not however solve the problem of deriving the value and uncertainty of their measurand X [14] as it can easily be shown by using an uncertainty representation by intervals. Then the value of X should certainly be inside the common part of the two intervals but the amount of uncertainty to be attributed to it is questionable; to simply apply its basic definition given by the GUM would lead to assume as uncertainty of X the common part of the uncertainty intervals of the two results. This would however mean to assume a lower uncertainty the smaller is this common part i.e. the higher is the difference between the results. To apply the GUM formulas to some weighted mean between m_1, m_2 would at contrary lead to uncertainties independent on this difference.

The PD to be considered for the difference $d=m_1-m_2$ between the outputs of two independent measurements, performed on the same measurand, may have different forms, according to the hypothesis that have been assumed for the PDs in the uncertainty interval (usually Gaussian or uniform). It has however always zero mean value and a SD s_d given by formula (13). From this distribution, an interval can be found in which, with an appropriate confidence level (e.g., 95%), the hypothesis of a unique value of X could not be rejected.

The opposite is however not true: to have two results which satisfy the condition expressed by the formula (10) does not necessarily correspond to the presence of a unique, unknown measurand X, even when its intrinsic uncertainty may be considered as negligible. When performing tests in a laboratory, this dilemma could be solved by increasing the number of measurements, but this is not always the case for the comparison of results in industrial tests.

When a non negligible value of intrinsic uncertainty u(X) is attributed to the measurand, each result has a larger uncertainty, then the PD of the quantity d, resulting from the combination of the PD of the two measurement results, is more complex and has a larger SD. Compatibility criteria should therefore be relaxed so that, e.g., the condition expressed by the formula (10) should become:

$$| m_1 - m_2 | \leq u_1 + u_2 + 2\, u(\mathrm{X}) \qquad (14)$$

Note however that, in this case, compatible measurement results may have no common part so that a compatibility condition based on the formula (10) becomes here in general non-transitive.

3.4. Further applications of compatibility

Compatibility may have, in principle, several applications that have not yet been fully explored.

A simple example is to use uncertainty intervals in determining the parameters of a linear regression between two measured quantities. The consequences of this approach would differ from those derived by the traditional use of least squares methods. For instance, a range of values (i.e., an interval of uncertainty) could be already assigned to the two parameters of a straight line when considering only two observations, but this range could even disappear when more observations are taken into consideration, if they are not compatible with a linear regression.

3.5. Tolerance limits

The general term “tolerance” indicates acceptable limits around a specified value, but it can have in practice two different meanings. Failing an internationally adopted terminology, the following distinction can be made [15]:

- when considering the construction of a part of a system, there are limits within which the values of its parameters may be considered as equivalent with respect to the system performance; the classical example is the tolerance on the rated values for the diameters of a cylindrical rod and its bearing, in connection with their good coupling; these limits may therefore be denoted as “constructive”;
- when considering the acceptance test for a device, the limits should be specified within which the measurement results must be included to assume that the device has passed the test; these limits have a more legal sense and are often expressed as “inspection limits”.

Clearly, the two meanings would coincide if uncertainty would not be present or not taken into account. This is generally the case when uncertainty is much lower than tolerance (e.g., in a ratio lower than 1:5). In any other case, different points of view may exist: a tolerance and an uncertainty may be simply added, to consider their worst possible combination; or they may be treated as a combination of two random variables with known PDs. For a quantitative assessment, with the introduction of uncertainty it becomes possible either to start from given constructive limits and impose lower inspection limits, or to reduce the constructive limits. Both these decisions however generally involve a cost as they require more sophisticated test and measuring equipment or higher quality for the apparatus to be manufactured.

It is also possible to subdivide the inspection limits into a zone of full conformance with specifications and zones of ambiguity, where neither conformance or non conformance can be proved, because the result is within the uncertainty intervals around the specified constructive limits. To manage this state of ambiguity and to reach a decision on the acceptance or rejection of a part it is necessary either to reduce uncertainties, by improving the measurements and the test procedure, or to accept the risk of a wrong statement on the conformance or non-conformance with specifications.

4. CONCLUSIONS

In the present paper it has been shown how the non-exactness of measurement results can be interpreted in terms of non-specificity and uncertainty, and how the latter can be defined, evaluated and expressed according to an international agreement, as formalized in the “Guide to the expression of uncertainty in measurement”. Such an agreement is specifically satisfactory for instrument calibrations, but it requires some extensions to be applied to other kinds of measurement.

The procedures described in the Guide could be implemented into an automatic measuring system, to directly obtain the measurement results in their complete form, i.e., as couples <measurand value, uncertainty estimate on it>. This would be particularly valuable for many types of industrial measurements, where the nature and characteristics of the measurands, of the concerned instruments, and of their influence quantities are well known and repeatable. Whenever the above mentioned conditions are not fulfilled, the evaluation of uncertainty remains a task of the operator, who must resort to his skill and experience to assess the range of values that can be reasonably attributed to the measurand, to dissipate the cloud of doubt that always exists when a physical phenomenon is expressed in some linguistic form. In these cases the capacity of critically analyzing and interpreting facts and instrument readings remain an exclusiveness of the human mind.

As the great Italian artist and scientist Leonardo da Vinci stated: «to have few thoughts involves to have many mistakes». This is one of the reasons why to measure is always a fascinating adventure.

REFERENCES

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

[2] I.Hacking, “Representing and intervening - Introductory topics in the philosophy of natural science”, Cambridge University Press, 1983.
[3] L.Mari, “The role of determination and assignment in measurement”, Measurement, 21 (3) (1997) 79-90.
[4] P.W.Bridgman, “How much rigor is possible in physics?”, in: Henkin et al. (eds.), The Axiomatic method, North-Holland, 1959.
[5] L.Mari, “On the comparison of inexact measurement results”, Measurement, 9 (4) (1991) 157-162.
[6] D.Krantz, R.Luce, P.Suppes, A.Tversky, “Foundations of measurement”, Academic Press, New York, Vol.1 (1971), Vol.2 (1989), Vol.3 (1990).
[7] EAL, “Expression of the uncertainty of measurement in calibration”, Publication EAL-R2, 1997.
[8] K.Weise, W.Woeger, “A Bayesian theory of measurement uncertainty, Meas. Sci. Techn., 4 (1993) 1-11.
[9] E.Arri, G.Iuculano, A.Zanobini, “Bayesian approach to uncertainty expression in a single measurement”, XIII IMEKO World Congress, Torino, 1994.
[10] ISO, IEC, “International vocabulary of basic and general terms in metrology”, 1993.
[11] K.Weise, W.Woeger, “Comparison of two measurement results using the Bayesian theory of measurement uncertainty”, Meas. Science Techn. 5 (1994) 879-882.
[12] S.D’Emilio, F.Galliana, E.Arri, G.Iuculano, “Compatibility of measurement results within measurement systems using hypotheses tests”, VIII Congrès Int. de Metrologie, Besancon, Oct 1997, pp.589-594.
[13] E.Arri, F.Cabiati, S.D’Emilio, “Traceability of electrical instruments and compatibility of measurement results”, 7th Symposium of IMEKO TC 4, Prague, 1995, pp.58-62.
[14] G.Zingales, “Compatibility in industrial measurements”, Paper to be presented at IMTC99, Venice, June 1999.
[15] G.Zingales. “Uncertainties and tolerances in H.V. tests”, Proceedings 10th Int. Symp. on High Voltage Engineering, Montreal, 1997, vol.4, pp.59-62.

Contact point: *Luca Mari*, Libero Istituto Universitario C.Cattaneo,
C.so Matteotti, 22, 21053 Castellanza (Va), ITALY
Phone: +39 331 572 1 Fax: +39 331 572320
E-mail: lmari@liuc.it