Metrologia 44 (2007) 513–529 doi:10.1088/0026-1394/44/6/011

Evolution of modern approaches to express uncertainty in measurement

Raghu Kacker[1], **Klaus-Dieter Sommer**[2] **and Rüdiger Kessel**[3]

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA
[2] Physikalisch-Technische Bundesanstalt, D-38116 Braunschweig, Germany
[3] Metrodata GmbH, D-79576 Weil am Rhein, Germany

E-mail: raghu.kacker@nist.gov

Received 11 April 2007
Published 27 November 2007
Online at stacks.iop.org/Met/44/513

Abstract
An object of this paper is to discuss the logical development of the concept of uncertainty in measurement and the methods for its quantification from the classical error analysis to the modern approaches based on the *Guide to the Expression of Uncertainty in Measurement (GUM)*. We review authoritative literature on error analysis and then discuss its limitations which motivated the experts from the International Committee for Weights and Measures (CIPM), the International Bureau of Weights and Measures (BIPM) and various national metrology institutes to develop specific recommendations which form the basis of the GUM. We discuss the new concepts introduced by the GUM and their merits and limitations. The limitations of the GUM led the BIPM Joint Committee on Guides in Metrology to develop an alternative approach—the draft Supplement 1 to the GUM (draft GUM-S1). We discuss the draft GUM-S1 and its merits and limitations. We hope this discussion will lead to a more effective use of the GUM and the draft GUM-S1 and stimulate investigations leading to further improvements in the methods to quantify uncertainty in measurement.

1. Introduction

Most metrologists are familiar with the rudimentary concepts of error analysis, recognize the propagation of uncertainties formula and are aware of the concepts of Type A and Type B evaluations promulgated by the 1995 *Guide to the Expression of Uncertainty in Measurement (GUM)* [1][4]. Many metrologists have heard of the draft Supplement 1 to the GUM (draft GUM-S1) [2] on propagation of probability density functions (pdfs) by numerical simulation. Some discussions and presentations on these approaches in conferences and meetings indicate inadequate understanding of the underlying probabilistic, statistical and metrological concepts and how these concepts depart from classical error analysis. In this paper we discuss why and how the concepts that underlie the GUM and the draft GUM-S1 were developed. We believe that knowledge of this background would enhance understanding, improve teaching and promote effective use of the GUM and the draft GUM-S1.

We review authoritative literature on the statistical concepts that underlie error analysis. The limitations of error analysis were a hindrance to communication of scientific and technical measurements. So leading authorities in metrology assembled, discussed and debated the issues in the late 1970s and came up with specific recommendations which form the basis of the GUM [1, Introduction]. A brief history of these developments is summarized in [1, Foreword], [3] and [4]. We discuss the logical development of the concept of uncertainty in measurement from error analysis to the GUM. Then we discuss the new concepts introduced in the GUM, which depart substantially from the preceding traditions of error analysis. Next we discuss the merits and limitations of the GUM. The limitations of the GUM motivated the BIPM Joint Committee on Guides in Metrology (JCGM) to develop an alternative made possible by recent advances in

[4] The GUM is published by the International Organization for Standardization (ISO) in the names of seven international scientific organizations: International Bureau of Weights Measures (BIPM), International Electro-technical Commission (IEC), International Federation of Clinical Chemistry (IFCC), International Organization for Standardization (ISO), International Union of Pure and Applied Chemistry (IUPAC), International Union of Pure and Applied Physics (IUPAP) and International Organization of Legal Metrology (OIML).

computer technologies. The alternative is the draft GUM-S1 on propagation of probability distributions by Monte Carlo simulation [2]. So, we review the draft GUM-S1 and discuss its merits and limitations. We hope this discussion will lead to a more effective use of the GUM and the draft GUM-S1 and stimulate further investigations to address the limitations of these approaches to quantify uncertainty in measurement.

1.1. Notation

We use Greek symbols such as τ, β, μ and σ for the values of metrological quantities and statistical parameters of probability distributions. We use lower-case Latin symbols such as x, y and w for data and their summary statistics. In statistical analysis, the data and their summaries are regarded as realizations of random variables having *sampling probability distributions*. A sampling probability distribution is a property of the data generation process (that is, the measurement procedure). Every sampling distribution has one or more unknown statistical parameters which are to be estimated. In both conventional and Bayesian statistical methods, the *likelihood function* of an unknown parameter is the sampling distribution of the data conditional on that parameter. We use the same lower-case Latin symbols such as x, y and w for random variables with sampling probability distributions and for their realized values. The context makes clear whether a symbol is a random variable or its realized value. As in the GUM, we use upper-case Latin symbols such as X, Y and W to represent random variables having *state-of-knowledge probability distributions* about the values of a quantity. A state-of-knowledge probability distribution represents belief probabilities about the possible values of a quantity based on all available information. All parameters of a state-of-knowledge probability distribution are fully specified. Common examples of statistical state-of-knowledge probability distributions are Bayesian prior and posterior distributions. A non-statistical example is a rectangular probability distribution for an unknown quantity specified by scientific judgment.

2. Review of error analysis

This review of the main points of error analysis is based on [4–15].

2.1. Authoritative references on error analysis in metrology

A classical approach to quantify uncertainty in measurement is error analysis. It is based on statistical sampling theory, which is also known as frequentist statistics, classical statistics or conventional statistics. The most authoritative references on error analysis in metrology include [5–8]. The popular books on error analysis include [9, 10]. Other important references relating to error analysis include [4, 11–15]. Reference [15] describes the logical growth of the concept of uncertainty from error analysis. However [15] is not easily accessible; therefore, we have included it as an appendix of this paper. The primary author of [15], Churchill Eisenhart, was a pioneer in developing conventional statistical methods for measurement science and technology and a leading authority on error analysis. In 1947, he founded the Statistical Engineering Division of the US National Bureau of Standards (NBS)[5].

2.2. True value, result of measurement and error

In error analysis, the quantity to be measured (a property of matter or phenomenon) is hypothesized (assumed) to have an unknown constant value (essentially unique and stable value) called *true value* denoted here by τ. A *result of measurement* for τ is an estimate denoted here by x. When it is useful to indicate the relationship between an estimate x and the corresponding parameter τ, we denote the latter by $\tau(x)$. The difference between x and τ is *error*[6]. In an ordinary measurement, x is the arithmetic mean of a series of n replicate measurements $q_1, \ldots, q_n$ made to estimate τ. The integer n may be one or more; in particular, the individual measurements are also results of measurement for τ. *True value and error are unknowable quantities* [1, annex D], except in the trivial case where τ is a finite number of entities.

2.3. Importance of statistical control of the measurement procedure

A simple error analysis is based on the assumption that all possible measurements that could be made under the given (assumed to be fixed) conditions form a fixed probability distribution and that the current measurements $q_1, \ldots, q_n$ may be regarded as a *random sample* from that distribution. Statistically this assumptions means that the measurement data $q_1, \ldots, q_n$ may be regarded as realizations of independent random variables (also denoted by $q_1, \ldots, q_n$) that have the same fixed sampling probability distribution with some expected value μ and some variance σ^2. This statistical assumption can be attributed to the data $q_1, \ldots, q_n$ only if the measurement procedure is in a state of statistical control [5, sections 3 and 4]. A state of statistical control is often demonstrated by periodically measuring a reference artefact of stable value for an extended period of time. The stability of the reference artefact may be confirmed by a more precise method. The importance of statistical control[7] of the measurement procedure cannot be overstated. We quote Eisenhart [5, section 4.1].

> 'In the foregoing we have stressed that a measurement operation to qualify as a measurement process must have attained a state of statistical control; and that until a measurement operation has been 'debugged' to the extent that it has attained a state of statistical control, it cannot be regarded in any logical sense as measuring anything at all.'

The form of the sampling distribution of $q_1, \ldots, q_n$ is often assumed to be approximately normal (Gaussian) with the

[5] NBS is the earlier name of the National Institute of Standards and Technology (NIST), the national metrology institute (NMI) of the USA.

[6] It is sometimes more appropriate to define error as x/τ, called fractional error. The concepts discussed in this paper can be extended to fractional error; however, we have not done that.

[7] Strict statistical control is extremely rare; the mean of the measurement process may fluctuate. In that case the variance σ^2 should include a component of variance for the fluctuation of process mean. If the process variance is unstable then the process is not in a state of statistical control.

expected value μ and variance σ^2. The assumption of normal distribution is utopian and it may be impractical or difficult to realize and maintain for an extended period of time.

It is important to be definite on what constitutes replicate measurements $q_1, \ldots, q_n$ because the extent to which conditions of measurement are allowed to vary freely over successive repetitions determines the scope of statistical inferences that may be drawn from the replicate measurements [5, section 4.1]. When the conditions of measurement are allowed to vary widely, the assumption of a fixed normal distribution may be doubtful.

2.4. Sampling probability distribution of the mean

When the measurement procedure is in a state of statistical control, the result x (arithmetic mean of the n measurements $q_1, \ldots, q_n$) may be regarded as a realization of a random variable (also denoted by x) with a fixed *sampling probability distribution*[8] with the expected value $E(x) = \mu$ and variance $V(x) = \sigma^2(x) = \sigma^2/n$. The claim that the variance $\sigma^2(x)$ is equal to σ^2/n requires that the measurements $q_1, \ldots, q_n$ be mutually independent[9]. The variance $\sigma^2(x) = \sigma^2/n$ of the mean x tends to zero as the number n of measurements tends to infinity. Therefore, the expected value μ is the limiting value of the mean x as n tends to infinity. If the common sampling distribution of $q_1, \ldots, q_n$ is normal then the sampling distribution of x is also normal. When the measurement procedure is in statistical control and n is not too small, the sampling probability distribution of the mean x may be approximately normal even when the probability distributions of $q_1, \ldots, q_n$ are not normal. The latter result is based on the *central limit theorem* [13]. Certain statistical methods require that the original measurements $q_1, \ldots, q_n$ must be normally distributed.

2.5. Random error and bias

Error analysis is based on parsing the error $x - \tau$ into two parts: $(x - \tau) = (x - \mu) + (\mu - \tau) = e(x) + \beta$, where $e(x) = x - \mu$ is *random error* and $\beta = \mu - \tau$ is *systematic error*, also called *bias* or *offset*. Parsing of error in this way amounts to postulating the following *statistical model* for the result x:

$$x = \tau + \beta + e(x). \tag{1}$$

In model (1), neither the true value τ nor the bias β can ever be known exactly.

2.6. Precision and accuracy

The closeness of the mutual agreement between independent results of measurement is called *precision* and the closeness of the agreement between independent results of measurement and the true value is called *accuracy* [5]. Precision and accuracy are characteristics of the measurement procedure employed and not of a particular result obtained [6]. That is, precision and accuracy are characteristics of the sampling probability distribution of measurements rather than of a realized result. When the measurement procedure is in a state of statistical control, the variance σ^2 is a measure of the precision of the sampling probability distribution of the measurements. As σ^2 decreases, precision increases; therefore, σ^2 is actually a measure of the *imprecision* [5], where imprecision is the opposite of precision. Likewise, inaccuracy is the opposite of accuracy.

2.7. Estimate of the variance

Suppose $s^2 = \sum_i (q_i - x)^2/(n-1)$ is the sampling theory estimate of the variance σ^2 determined from the current measurements $q_1, \ldots, q_n$. Then $s^2(x) = s^2/n$ is an estimate of the variance $\sigma^2(x) = \sigma^2/n$ based on the current measurements $q_1, \ldots, q_n$. If the measurement procedure is in a state of statistical control (which it must be), then one may have a better value for σ^2 based on large amounts of past data than the estimate s^2 based on the current measurements only. In that case the better value for σ^2 should be used.

2.8. Confidence intervals

Suppose $q_1, \ldots, q_n$ have the same normal sampling distribution, $N(\mu, \sigma^2)$, and are mutually independent. Then a confidence interval for the unknown expected value μ such as $(x \pm t_{(1-\alpha/2)} \times s/\sqrt{n})$, where $t_{(1-\alpha/2)}$ is $(1 - \alpha/2) \times 100$th percentile of the t-distribution with degrees of freedom $n - 1$, is a random interval. The location of the interval $(x \pm t_{(1-\alpha/2)} \times s/\sqrt{n})$ is random because x is random. The width of the interval $(x \pm t_{(1-\alpha/2)} \times s/\sqrt{n})$ is random because s^2 is random. The confidence level associated with $(x \pm t_{(1-\alpha/2)} \times s/\sqrt{n})$ is a statement about the sampling distributions of x and s^2 determined from $q_1, \ldots, q_n$. If the process of making the n measurements $q_1, \ldots, q_n$ could be repeated infinitely many times and the assumed fixed normal sampling distribution for $q_1, \ldots, q_n$ continued to apply then the fraction $(1-\alpha)$ of such intervals would include the unknown value μ. A computed confidence interval is a realization of the random interval $(x \pm t_{(1-\alpha/2)} \times s/\sqrt{n})$. The confidence level is not a statement of probability concerning the computed interval.

A confidence interval for σ^2 is a random interval of the form $[(n-1)s^2/a, (n-1)s^2/b]$, where a and b are suitably chosen percentiles of the chi-square distribution with degrees of freedom $n - 1$. If the process of making measurements could be repeated infinitely many times then some fraction (determined by the choices of a and b) of such intervals would include the unknown value σ^2. The confidence level is not a statement of probability concerning the computed confidence interval.

2.9. Bound on bias

The bias β in the measurements $q_1, \ldots, q_n$ is the unknown constant parameter $\mu - \tau$, where μ is the expected value of the sampling distributions of the measurements and τ is the true value of the measurand. The measurements $q_1, \ldots, q_n$ carry

[8] Sampling probability distribution is equally important in conventional statistics and Bayesian statistics. In Bayesian statistics, sampling distribution is used to define the likelihood function.

[9] It is easy to show (using propagation of variances formula) that if the correlation coefficient between q_i and q_j is ρ_{ij}, for $i \neq j$, then $V(x) = [\sigma^2/n][1 + (n-1)\rho]$, where ρ is the average correlation coefficient $\rho = \Sigma\rho_{ij}/(n(n-1))$ of the $n(n-1)$ pairs $\{q_i, q_j\}$ for $i \neq j$. When the average correlation coefficient ρ is positive then the expression σ^2/n understates the variance $V(x)$.

no information about their bias. In error analysis, a bound δ on the possible bias (offset) in the measurements is specified by expert judgment such that the probability that the bias β may exceed the limits of the interval $(-\delta, \delta)$ is believed to be zero. That is, $-\delta \leqslant \beta \leqslant \delta$ with probability one.

2.10. Uncertainty

In error analysis, the *uncertainty* of the sampling distribution of a result x is indicated by stating credible limits to its likely inaccuracy [6]. The inaccuracy is indicated by two expressions: (i) an estimate of the imprecision of x, such as $s^2(x)$, and (ii) an assessment of the bound δ on its bias β.

2.11. Propagation of errors

Often, a result of measurement for the quantity of interest cannot be determined by the direct measurement of that quantity, but is determined from measurements of one or more other quantities through a known functional relationship. For example, generally the speed v of a moving object is not measured directly but is determined from the measurements of the time t taken to cover a distance d through the relationship $v = d/t$. In such cases an estimate of imprecision and an assessment of the bound on bias for the output quantity of interest (such as speed) are determined from the estimates of imprecision and the assessments of the bounds on biases for the input quantities (such as the distance and the time).

Suppose a result of measurement w is determined from the two results of measurement x_1, x_2 through a given function $w = g(x_1 x_2)$. Suppose the unknown true values for the three quantities are $\tau(w)$, $\tau(x_1)$ and $\tau(x_2)$, respectively. The function $w = g(x_1, x_2)$ is based on scientific theories and empirical knowledge which imply that $\tau(w) = g(\tau(x_1), \tau(x_2))$. Suppose the expected value and variance of the sampling probability distribution of x_i are $\mu(x_i)$ and $\sigma^2(x_i)$, for $i = 1, 2$ and the expected value and variance of the corresponding w are $\mu(w)$ and $\sigma^2(w)$, respectively. For simplicity assume that the sampling distributions of x_1 and x_2 are mutually independent. A linear Taylor series approximation of the function $w = g(x_1, x_2)$ is

$$w = g(x_1, x_2) \approx g(\mu(x_1), \mu(x_2)) + c_1(x_1 - \mu(x_1)) + c_2(x_2 - \mu(x_2)), \tag{2}$$

where the constant coefficients c_1 and c_2, called sensitivity coefficients, are partial derivatives of the function g evaluated at $\mu(x_1)$ and $\mu(x_2)$. The expected value and variance of (2) are, respectively,

$$E(w) = \mu(w) \approx g(\mu(x_1), \mu(x_2)) \tag{3}$$

and

$$V(w) = \sigma^2(w) \approx c_1^2\sigma^2(x_1) + c_2^2\sigma^2(x_2). \tag{4}$$

Note that the expected value $\mu(w)$ of w is not $g(\mu(x_1), \mu(x_2))$ unless the function $g(x_1, x_2)$ is linear, in which case expression (2) is not an approximation. When the expected values ($\mu(x_1)$ and $\mu(x_2)$) and variances ($\sigma^2(x_1)$ and $\sigma^2(x_2)$) are unknown, they are replaced in (4) with their estimates (x_1 and x_2) and ($s^2(x_1)$ and $s^2(x_2)$) to get the following weaker approximation than (4):

$$s^2(w) \approx c_1^2 s^2(x_1) + c_2^2 s^2(x_2)\ , \tag{5}$$

where the sensitivity coefficients c_1 and c_2 are now evaluated at the known estimates x_1 and x_2.

Expression (5) is traditionally called *law of error propagation*. This is a misnomer. A linear approximation for the errors is $w \approx \tau(w) + c_1(x_1 - \tau(x_1)) + c_2(x_2 - \tau(x_2))$. This expression does not lead to (4) and then to (5) because the expected values $E(x_i - \tau(x_i))^2$ are not equal to $\sigma^2(x_i)$, for $i = 1, 2$. The difference between $E(x_i - \tau(x_i))^2$ and $\sigma^2(x_i)$ is the square of the bias $\mu(x_i) - \tau(x_i)$ which is unknowable.

Suppose the bounds on biases in x_1 and x_2 are assessed as $\delta(x_1)$ and $\delta(x_2)$, respectively. Then one of the many ways of approximating the bound on bias in w is

$$\delta(w) \approx |c_1\delta(x_1)| + |c_2\delta(x_2)|, \tag{6}$$

where the quantities on the right side of (6) are absolute values [8]. Expressions (5) and (6) can be extended to a function of more than two input quantities and for correlated results [8].

2.12. Report from error analysis

In error analysis, an estimate of imprecision and an assessment of the bound on bias are different components of inaccuracy and they are subject to different treatments in usage. Therefore they are stated separately. We quote Eisenhart [5, section 4.3].

> 'By whatever means credible bounds to the likely overall systematic error of the measurement process are obtained they should not be combined (by simple addition, by quadrature, or otherwise) with the experimentally determined measure of its standard deviation to obtain an overall index of its accuracy (or more correctly of inaccuracy). Rather (a) the standard deviation of the process and (b) credible bounds to its systematic error should be stated separately.'

Subsequently, Eisenhart [6] relaxed the above recommendation as follows.

> 'The above recommendation should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on the inaccuracy, that is, on the overall uncertainty, of a reported value, provided that separate statements of its imprecision and its possible systematic error are included also.'

2.13. Uncertainty associated with reference standards

In error analysis, the uncertainties associated with the accepted values of reference standards and the values assigned to the fundamental constants of nature are not ordinarily included in determining the bounds on bias for the measurement procedure. The reason given in [6] is that their inclusion would make everybody's results appear unduly inaccurate relative to each other. The GUM [1, annex D] departs from this viewpoint.

3. Evolution of the concept of uncertainty from error analysis to the GUM

In this section, we discuss logical development of the concept of uncertainty in measurement from error analysis to the GUM.

3.1. Error and uncertainty in measurement

The ideas of error and uncertainty were mixed up before publication of the GUM. The two terms were used interchangeably [10, p 3], [14, p 241] or uncertainty was used to refer to an estimate of error [9, p 6] or an estimate of the likely limits of error [15]. The GUM [1, annex D] makes clear that error is an unknowable quantity in the realm of the *state of nature* and uncertainty (in measurement) is a quantifiable parameter in the realm of the *state of knowledge about nature*. Uncertainty represents the dispersion of the values that could reasonably be attributed to the measurand [1, annex B]. A stated result of measurement is one value (albeit a central one) that could be attributed to the measurand. The result of measurement and its associated uncertainty together represent a range of the values that could be attributed to the measurand with varying degrees of credibility. Uncertainty is not an estimate of the likely limits of error because, for example, a systematic effect might have been overlooked because it is unrecognized [1, section D.5.1]. Only when there is a sound basis to believe that no systematic effects have been overlooked uncertainty may be considered as an approximate measure of possible error [1, section D.6.1]. A result of measurement and uncertainty are determined from all available information. The GUM method of evaluating uncertainty is transparent and it is clear what effects are included in the combined uncertainty.

3.2. Overall expression of uncertainty

This section is a highly simplified discussion of a main point from [15]. As science and technology advanced, publications started including tables of measurement results from various sources. This generated the need for an overall expression of uncertainty for the tabulated results, which raised the question, how could a sampling theory estimate of imprecision (such as $s^2(x)$) and a judgmental bound on bias (such as δ) be logically combined. Several methods were developed. An elementary method was to add the bound δ to the estimate $s(x)$ resulting in the following expression:

$$u(x) = s(x) + \delta, \tag{7}$$

for the overall uncertainty. Since the estimate $s^2(x)$ carries statistical uncertainty indicated by its degrees of freedom, it was argued that $s(x)$ and δ should be added at a similar confidence level [15]. Thus, for normally distributed measurements the following expression for the overall uncertainty $u(x)$ was developed:

$$u(x) = t_\nu(\alpha) \times s(x) + \delta, \tag{8}$$

where $s^2(x)$ is based on ν degrees of freedom and $t_\nu(\alpha)$ is a percentile of Student's t-distribution for confidence level α. A third method was to add $s(x)$ and δ in the quadrature, resulting in the expression

$$u(x) = \sqrt{s^2(x) + \delta^2}. \tag{9}$$

If the unknown constant bias β were regarded as a random variable with two possible values $-\delta$ and $+\delta$ each with probability $\frac{1}{2}$ then the expected value of β would be zero and the variance of β would be δ^2. In the statistical model (1), $x = \tau + e(x) + \beta$, the variance of the sampling distribution of $e(x)$ is $\sigma^2(x)$ with estimate $s^2(x) = s^2/n$; therefore (9) may be regarded as an estimate of the standard deviation of x. The idea that the bias β may be regarded as a random variable led to the idea of assigning to the bias β a state-of-knowledge rectangular distribution on the interval $(-\delta, \delta)$ with variance $\delta^2/3$. Thus the square δ^2 of the bound on bias in (9) was replaced with the variance $\delta^2/3$ resulting in the following expression for the overall uncertainty

$$u(x) = k\sqrt{s^2(x) + \delta^2/3}, \tag{10}$$

where k is customarily taken as 2 or 3. Expression (10) was proposed in [16] and used in the PTB[10]. It was termed the PTB approach [15]. This approach is a precursor to the GUM. Expression (10) needed a probabilistic interpretation. In part motivated by this need, the GUM developed the concepts of Type A and Type B evaluations of uncertainty and the concept of measurement equation.

3.3. Whose uncertainty: metrologist or user

In a report from error analysis, the uncertainty of a result of measurement x is expressed in different forms depending on the importance of the estimate of imprecision and the assessment of the bound on bias in relation to the intended use of the result x as well as to other possible uses to which it may be put [6, 7, 13, chapter 23]. Thus the following four cases are considered: (i) both bias and imprecision are negligible, (ii) bias is not negligible but imprecision is negligible, (iii) neither bias nor imprecision is negligible and (iv) bias is negligible but imprecision is not negligible. Specific recommendations with respect to each of these cases are discussed in [6]. A user is expected to use the report from error analysis to determine his/her expression of uncertainty in light of the intended use of measurements. Thus error analysis puts the responsibility for evaluating uncertainties on the users of the measurement data.

The GUM puts the responsibility for quantifying uncertainty on the metrologist who makes the measurements. The stated result of measurement and its associated uncertainty together indicate the state of knowledge of the metrologist concerning the unknown value of the quantity measured rather than the uncertainty associated with a subsequent use of the result of measurement.

3.4. Random and systematic sources of variation

A measurement procedure is affected by various sources of variation and it is often difficult to determine the category, random or systematic, to which a source of variation should be assigned [1, section E.1.3], [3, 14, 15]. The decision depends on the chosen viewpoint (interpretation) of the measurement procedure. Thus arbitrariness is inherent in the classification of errors as random and systematic. Despite arbitrariness of this classification, in error analysis an estimate of imprecision (arising from random error) and an assessment of the bound on bias (systematic error) are reported separately or combined in an ad hoc manner. The GUM regards the variations

[10] PTB (Physikalisch-Technische Bundesanstalt) is the NMI of Germany.

arising from random and systematic effects as components of uncertainty and recommends a logically consistent way of combining all components of uncertainty.

3.5. Importance of realistic uncertainty

In error analysis, a bound on bias is set to a safe value that is not likely to be exceeded. This often made the uncertainty unrealistically large [1, section E.1]. When the uncertainty is generally stated to be too large, decision makers discount it or ignore it completely. Indeed, catastrophes have happened because stated uncertainties were ignored. On the other hand, an understatement of uncertainty might lead to misplaced trust on the reported values. For example, when the stated uncertainty is too small, an inspector may reject with higher frequency manufactured products that are in-tolerance and accept with higher frequency those that are out-of-tolerance. The GUM [1, annex E] emphasizes the importance of determining realistic uncertainties to the extent of the state of knowledge of metrologist.

3.6. Target value of the measurand

Eisenhart [5] discusses at length the concept of true value in error analysis and the difficulty of defining it; we quote from [5, section 3.3c].

> 'Indeed, as is evident from the foregoing, the 'true value' of the magnitude of a particular quantity is intimately linked to the purposes for which a value of the magnitude of this quantity is needed, and its 'true value' cannot, in the final analysis, be defined meaningfully and usefully in isolation from these needs. Therefore, as this fact becomes more widely recognized in science and engineering, I hope that the traditional term 'true value' will be discarded in measurement theory and practice, and replaced by some more appropriate term such as '*target value*' that conveys the idea of being the value that one would like to obtain for the purpose in hand, without any implication that it is some sort of permanent constant preexisting and transcending any use that we may have for it.'

The GUM [1, annex D] discusses the concept of true value at length. It states that depending on the detail with which the measurand is defined, a range of values may be consistent with the definition of measurand and the adjective 'true' in true value is unnecessary.

We think that a modifier for the word value when referring to a value of the measurand may be useful for ease of communication. For the reasons discussed in [5, section 3.3c], we prefer the term target value over the term true value when referring to a value of the measurand.

4. *Guide to the Expression of Uncertainty in Measurement* (GUM)

In this section we discuss statistical and metrological concepts that underlie the GUM. The draft GUM-S1 is based on certain variations of these concepts which are discussed in the next section.

4.1. Measurement equation in the GUM

The concept of measurement equation is the most original methodological contribution of the GUM, in our view. The term measurement equation[11] was introduced in [2] by two of the primary authors of the GUM (Dr Barry N Taylor and Dr Chris E Kuyatt of NIST). In the GUM, the measurement equation is a function

$$Y = f(X_1, ..., X_N) \tag{11}$$

that represents the after-measurement-method (ingredients and recipe) of determining a result y for the value of the measurand and its associated standard uncertainty $u(y)$ from various input values $x_1, \ldots, x_N$ and their associated standard uncertainties $u(x_1), \ldots, u(x_N)$ and correlation coefficients $r(x_1, x_2), \ldots, r(x_{N-1}, x_N)$. The GUM regards all input arguments and the output of a measurement equation $Y = f(X_1, \ldots, X_N)$ as variables with state-of-knowledge probability distributions about the corresponding input and output quantities. Thus Y is a variable having a state-of-knowledge probability distribution about the value of the measurand and $X_1, \ldots, X_N$ are input variables having state-of-knowledge probability distributions about various input quantities. The input values $x_1, \ldots, x_N$ are identified with the expected values, the standard uncertainties $u(x_1), \ldots, u(x_N)$ are identified with the standard deviations and $r(x_1, x_2), \ldots, r(x_{N-1}, x_N)$ are correlation coefficients of the state-of-knowledge probability distributions for $X_1, \ldots, X_N$. Some of the input variables $X_1, \ldots, X_N$ may themselves be regarded as measurands and functions of additional variables. Thus the function $Y = f(X_1, \ldots, X_N)$ may represent a hierarchical system of equations [17, 18]. A measurement equation should not be confused with a statistical model[12] stipulated to relate the measurement data and statistical parameters.

Development of a thorough measurement equation is the key to determining uncertainty in measurement. To develop a measurement equation, detailed understanding of the important influence quantities in the measurement procedure is required [19]. All other aspects of uncertainty evaluation can be programmed in a computer.

In the GUM, the inputs are expected values, variances (squares of standard uncertainties) and correlation coefficients associated with state-of-knowledge probability distributions for the input quantities and the outputs are the expected value and variance for the measurand. Thus the inputs and the outputs of a measurement equation in the GUM are similar entities (such is not the case in the draft GUM-S1).

4.2. Type A and Type B evaluations

The uncertainty associated with a result of measurement generally consists of many components. The GUM explicitly

[11] It is unfortunate that the draft GUM-S1 uses the term 'measurement model' or 'model of measurement' or simply 'model' for a measurement equation because metrologists use the term measurement model for two different things: (i) a statistical model, such as (1), which relates the data to statistical parameters and (ii) the measurement equation, such as (11). Thus the term measurement model is highly ambiguous.

[12] Statistical models relate data to statistical parameters. For example, the one-way analysis of variance model $y_{ij} = \mu + \tau_i + e_{ij}$ relates the data $\{y_{ij}\}$ to the statistical parameters μ and $\{\tau_i\}$ and the simple linear regression model $y_i = \alpha + \beta x_i + e_i$ relates the data $\{y_i\}$ to the statistical parameters α and β and the values $\{x_i\}$ of the regressor variable.

recognizes non-statistical methods (termed Type B) as valid means to quantify components of uncertainty. Indeed, professionally determined Type B evaluations of uncertainty components may be more reliable than statistical evaluations (termed Type A) [1, section 4.3.2]. The terms Type A and Type B apply to the methods of evaluation rather than to the sources of uncertainty. Although the GUM [1, sections 2.3.2 and 2.3.3] applied the labels Type A and Type B to the methods for evaluating uncertainties, this classification also applies to the methods for evaluating expected values, standard deviations and correlation coefficients of state-of-knowledge probability distributions for the input variables of a measurement equation [1, section 5.2.5, 18].

Type B evaluations are parameters of state-of-knowledge probability distributions specified by judgment based on all available information [1, section 4.3]. All illustrations of Type A evaluations discussed in the GUM are estimates determined from frequentist statistics. Their probabilistic interpretation does not agree with state of knowledge probabilistic interpretation of Type B evaluations [18]; thus, they cannot be combined with Type B evaluations logically. The GUM [1, section 4.1.6] resolved this issue by declaring that the statistical estimates (Type A) determined from frequentist statistics be regarded as parameters of state-of-knowledge probability distributions. That is, all statistical (Type A) and non-statistical (Type B) evaluations in the GUM are regarded as parameters of state-of-knowledge probability distributions. Thus in the GUM, Type A evaluations, determined from frequentist statistics, acquired the same state-of-knowledge probabilistic interpretation as Type B evaluations. Armed with this interpretation Type A and Type B evaluations can be logically combined. The method of combining Type A and Type B evaluations was formalized by the concept of measurement equation.

4.3. Combining uncertainties through a linear approximation of the measurement equation

A major focus of the GUM is the quantification of the uncertainty associated with a result of measurement that is determined from a number of input quantities through a measurement equation. The GUM propagates the estimates, standard uncertainties and correlation coefficients for various input quantities through a linear approximation of the measurement equation to determine an estimate and standard uncertainty for the value of the measurand. Specifically, the GUM approximates the measurement equation (11) by the linear function

$$Y \approx Y_{\mathrm{Linear}} = f(x_1, \ldots, x_N) + \sum_{i=1}^{N} c_i (X_i - x_i), \qquad (12)$$

determined from the first-order Taylor series expansion, where $x_1, \ldots, x_N$ are input values for $X_1, \ldots, X_N$; the coefficients $c_1, \ldots, c_N$, called *sensitivity coefficients*, are partial derivatives of Y with respect to $X_1, \ldots, X_N$ evaluated at $x_1, \ldots, x_N$. If we regard $x_1, \ldots, x_N$ as the expected values, $u(x_1), \ldots, u(x_N)$ as the standard deviations and $r(x_1, x_2), \ldots, r(x_{N-1}, x_N)$ as the correlation coefficients of state-of-knowledge pdfs for $X_1, \ldots, X_N$ then the expected value and the variance of Y_{Linear} defined in (12) yield the following expressions for y and $u(y)$:

$$y = E(Y_{\mathrm{Linear}}) = f(x_1, \ldots, x_N) \qquad (13)$$

and

$$u^2(y) = V(Y_{\mathrm{Linear}}) = \sum_{i=1}^{N} c_i^2 u^2(x_i) + 2\sum_{i<j}^{N} c_i c_j u(x_i) u(x_j) r(x_i, x_j). \qquad (14)$$

Expression (14) is called the propagation of uncertainties formula. The GUM regards the result y and the standard uncertainty $u(y)$ determined from (13) and (14) as an approximate expected value and standard deviation of a state-of-knowledge probability distribution for the value Y of the measurand. The result y and the standard uncertainty $u(y)$ differ from the unknown expected value $E(Y)$ and the standard deviation $S(Y)$ of $Y = f(X_1, \ldots, X_N)$ to the extent that the probability distribution of Y differs from the probability distribution of Y_{Linear}. The GUM [1, section E.5] discusses the relationship between the propagation of uncertainties formula (14) and a similar propagation of errors formula used in error analysis.

A proper evaluation of uncertainty requires that every effort must be made to identify correlated input variables and correlation coefficients between those input variables must be quantified and included in the calculation of combined standard uncertainty [20,21]. Indeed, failure to identify and incorporate correlation coefficients between input variables is a leading cause of unreasonable uncertainty evaluations.

When the input variables are uncorrelated, the propagation of uncertainties formula (14) reduces to

$$u^2(y) = \sum_{i=1}^{N} c_i^2 u^2(x_i). \qquad (15)$$

Thus for uncorrelated input variables, the products $c_1 u(x_1), \ldots, c_N u(x_N)$ are the *uncertainty contributions* from the components $u(x_1), \ldots, u(x_N)$ to the combined standard uncertainty $u(y)$ [22,23].

The GUM [1, section 5.1.2] notes that when the non-linearity of the measurement equation is significant, higher order terms from the Taylor series expansion must be included in (15). The higher order terms given in the GUM [1, section 5.1.2] require not only that the pdfs of the input variables $X_1, \ldots, X_N$ be independent but also that each has a normal distribution [18, section 5.3]. The latter requirement is highly restrictive and not stated in the GUM. Most metrologists do not use the higher order terms given in the GUM to determine combined standard uncertainty.

4.4. Degrees of freedom

A statistical estimate determined from sampling theory (frequentist statistics) is uncertain because of the limited number of measurements [1, section E.4.3]. Such uncertainty in a sampling theory estimate is called statistical uncertainty. The statistical uncertainty in the estimate $s^2(x) = s^2/n$ of the variance $\sigma^2(x) = \sigma^2/n$ of the mean x of independent and identically normally distributed measurements $q_1, \ldots, q_n$ is quantified by degrees of freedom. The degrees of freedom associated with $s^2(x)$ are $n - 1$. Even though the GUM interprets $s^2(x)$ as a Type A evaluation of the variance of a state-of-knowledge pdf, it recognizes the degrees of freedom $n - 1$ as a measure of its statistical uncertainty.

In addition, the GUM [1, section G.4.2] suggests that the concept of degrees of freedom may be used to represent subjective doubt about a non-statistical (Type B) evaluation of uncertainty. This suggestion is controversial and most metrologists ignore it [25, section 2, comment 3].

If measurement equation (11) is linear and $X_1, \ldots, X_N$ are independently distributed, the GUM [1, section G.4] recommends the use of the Welch–Satterthwaite formula to determine effective degrees of freedom associated with the combined standard uncertainty $u(y)$ determined from (15).

It is shown in [18, 25, 26] that if Bayesian statistics is used for the statistical (Type A) evaluations, then there is no need to count degrees of freedom. Bayesian estimates are never uncertain; in particular, they do not have statistical uncertainties.

4.5. *Quantification of uncertainty from systematic effects*

A common practice in metrology is to apply a correction to the result of the measurement when a significant systematic effect is recognized. A correction for a systematic effect is always uncertain. Before publication of the GUM, there was no generally accepted approach to account for the uncertainty associated with a correction for a systematic effect. The GUM established the following principle. A correction should be applied to a result of measurement for each recognized[13] significant systematic effect and the uncertainty associated with the correction should be quantified and included in the combined standard uncertainty associated with the corrected result. The correction is regarded as a random variable with a state-of-knowledge pdf [1, section 4.3], [17, section D.3]. A state-of-knowledge pdf for the correction variable is specified on the basis of all available information including scientific judgment and relevant data. The expected value of this state-of-knowledge distribution is the correction applied to the result and the standard deviation is the associated uncertainty. Often the correction applied to the result is zero but the associated uncertainty could be large [17, section D.3.1].

There is no direct correspondence between the classification of the methods of evaluation as Type A or Type B and the identification of uncertainties as arising from random effects or from the corrections for systematic effects [17, section D.2]. However, the components of uncertainty arising from random effects are usually evaluated by statistical methods (Type A) and the components of uncertainty arising from the corrections for recognized systematic effects are most often evaluated by other methods (Type B).

4.6. *Uncertainty from definition and realization of the measurand*

Whereas error analysis is concerned only with the inaccuracy of a measurement procedure, the GUM recognizes sources of uncertainty relating to the definition and the realization of the measurand as well [1, section 3.3.2 (a), (b), (c) and annex D]. In practice, the measurand is incompletely defined [1, section D.1]; thus, many values may agree with the stated definition of measurand. The incomplete definition of the measurand is a source of uncertainty. The quantity actually realized for measurement may not fully agree with the definition of the measurand [1, section D.2]. The difference between a value of the quantity realized for measurement and a target value of the measurand is another source of uncertainty. These potential sources of uncertainty are often, but not always, negligible [1, section D.5.3].

4.7. *Fundamental expression of uncertainty*

The authors of the GUM [1, section 0.4] determined that a suitable expression of uncertainty should satisfy the following criteria:

> '*Internally consistent*: it should be directly derivable from the components that contribute to it, as well as independent of how these components are grouped and of the decompositions of the components into subcomponents.'
>
> '*Transferable*: it should be possible to use directly the uncertainty evaluated for one result as a component in evaluating the uncertainty of another measurement in which the first result is used.'

A standard uncertainty is both internally consistent and transferable. Therefore, standard uncertainty is the fundamental expression of uncertainty in both the GUM as well as the draft GUM-S1.

4.8. *Expanded uncertainty interval and coverage probability*

Sometimes it is necessary to express uncertainty as an interval [1, section 6.1.2]. To meet this need, the GUM introduced the concept of expanded uncertainty. An expanded uncertainty U is obtained by multiplying the standard uncertainty $u(y)$ by a coverage factor k; that is $U = k\,u(y)$. The expanded uncertainty U defines an interval $[y \pm U] = [y \pm k\,u(y)]$ about the result of measurement y that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand [1, section 3.3.7]. The GUM did not assign a name to the interval $[y \pm U]$. However, Dr Barry N Taylor, a primary author of the GUM, refers to it as an expanded uncertainty interval [18, reference 5]. In the GUM, an expanded uncertainty interval is determined after the result y and its associated standard uncertainty $u(y)$ have been determined. Along with an expanded uncertainty interval $[y \pm U]$, the coverage factor k must be stated, so the standard uncertainty $u(y)$ may be recovered.

The coverage probability of an expanded uncertainty interval $[y \pm k\,u(y)]$ is the fraction covered by this interval of a state-of-knowledge probability distribution for Y, represented by y and $u(y)$. Since the GUM does not yield a state-of-knowledge probability distribution for Y, the coverage probability cannot be determined in general. Sometimes, it may be possible to state an approximate coverage probability, when the measurement equation is

[13] If we can think of a systematic effect then it is recognized. A recognized effect may or may not be significant.

Table 1. Sketch of an uncertainty budget for uncorrelated input quantities.

Quantity X_i	Result x_i	Standard uncertainty $u(x_i)$	Sensitivity coefficient c_i	Uncertainty contribution $c_i u(x_i)$
X_1	x_1	$u(x_1)$	c_1	$c_1 u(x_1)$
X_N	x_N	$u(x_N)$	c_N	$c_N u(x_N)$
Y	y			$u(y)$

linear, the pdfs for $X_1, \ldots, X_N$ are independently distributed and the requirements of the central limit theorem are met [1, annex G, 18, 26].

4.9. Uncertainty budget

An uncertainty budget is a table that lists all input quantities (or their identifiers such as X_i), the corresponding results (estimates) x_i, standard uncertainties $u(x_i)$, sensitivity coefficients c_i and the uncertainty contributions $c_i u(x_i)$ [22, 23]. The types of methods (Type A or Type B) used for evaluating the results and uncertainties are also indicated. Along with numerical values, the units of measurement should also be stated. A sketch of an uncertainty budget for the uncorrelated input variable is shown in table 1.

The GUM did not suggest any format for reporting the details on how the combined standard uncertainty was determined. A task force of the European Co-operation for Accreditation [22] developed the standardized format indicated in table 1 for presenting the details. An uncertainty budget is a vital output from the evaluation of combined standard uncertainty $u(y)$ for the following reasons. An uncertainty budget makes the calculation of combined standard uncertainty transparent. A user of the results from uncertainty evaluation may be interested in the degrees of contribution to the combined standard uncertainty $u(y)$ from each of its components $u(x_1), \ldots, u(x_N)$ [22, 23]. The Eurachem/Citac-Guide [23] displays bar-charts of the uncertainty contributions $c_i u(x_i)$, for $i = 1, \ldots, N$, to show the degrees of contribution. In particular, the uncertainty budget identifies the dominant components of combined standard uncertainty $u(y)$. A metrologist who is interested in understanding, managing or improving the measurement process needs the details provided by the uncertainty budget. In interlaboratory evaluations uncertainty budgets from individual laboratories are needed for detailed comparisons, especially when the results are inconsistent. An uncertainty budget is a critical aid in assessing a reported combined standard uncertainty $u(y)$.

4.10. Degrees of contribution from correlated inputs

This subsection describes a useful addition to the uncertainty budget for correlated input variables. (It is not a part of the main GUM procedure.) Reference [24] introduced a coefficient of contribution, $h(y, x_i)$, which quantifies the degree of contribution from $u(x_i)$ to $u(y)$. The coefficients of contribution add up to one so they may be expressed as a per cent. For uncorrelated input variables, [24] defines the coefficient of contribution, $h(y, x_i)$, as

$$h(y, x_i) = \left[\frac{c_i u(x_i)}{u(y)}\right]^2 . \tag{16}$$

The coefficient of contribution (16) is the square of the correlation coefficient between Y_{Linear} and X_i. A general expression for the coefficient of contribution, $h(y, x_i)$, that applies to both uncorrelated and correlated input variables is

$$h(y, x_i) = \left[\frac{c_i u(x_i)}{u(y)}\right] [r(y, x_i)] , \tag{17}$$

where $r(y, x_i)$ is the coefficient of correlation between Y_{Linear} and X_i defined as

$$r(y, x_i) = \sum_{j=1}^{N} \left[\frac{c_j u(x_j)}{u(y)}\right] [r(x_i, x_j)] . \tag{18}$$

General expression (17) reduces to the special case (16) for uncorrelated input variables. The degrees of contribution from $u(x_1), \ldots, u(x_N)$ to $u(y)$ are affected by correlations between the input variables [24]. Therefore, when some input variables are correlated it is useful to include in the uncertainty budget a column for the correlation coefficients $r(y, x_i)$ and a column for the coefficients of contribution $h(y, x_i)$ [24].

4.11. Merits of the GUM

(i) The GUM does not require complete knowledge of the state-of-knowledge pdfs for the input quantities and their joint pdf.

(ii) The GUM requires as inputs only the expected values, standard deviations and correlation coefficients of state-of-knowledge probability distributions for the input quantities. It yields as output an approximate expected value and standard deviation of a state-of-knowledge probability distribution for the value of the measurand. Thus GUM is applicable as an approximation (to the extent that a linear approximation to the measurement equation gives reasonable results) in those situations where reasonable estimates for the expected values, standard deviations and correlation coefficients for the input variables are available.

(iii) When measurement equation (11) is linear, the result y and the standard uncertainty $u(y)$ obtained from (13) and (14) are the correct expected value and standard deviation of Y for all state-of-knowledge probability distributions for the input variables $X_1, \ldots, X_N$ that have the expected values $x_1, \ldots, x_N$, standard deviations $u(x_1), \ldots, u(x_N)$ and correlation coefficients $r(x_1, x_2), \ldots, r(x_{N-1}, x_N)$ [25]. Thus, when the measurement equation is linear, the result y and the standard uncertainty $u(y)$ determined according to the GUM are *robust evaluations* that apply for all pdfs with the given parameters for the input quantities.

4.12. Limitations of the GUM

(i) When measurement equation (11) is non-linear and the uncertainties $u(x_1), \ldots, u(x_N)$ are not sufficiently small, the result y and the standard uncertainty $u(y)$ determined from (13) and (14), respectively, may be poor approximations to the expected value $E(Y)$ and the standard deviation $S(Y)$ of $Y = f(X_1, \ldots, X_N)$ [25].
(ii) When measurement equation (11) is non-linear, the GUM does not provide any indication of the direction and magnitude of errors possible in the result y and the standard uncertainty $u(y)$ with respect to $E(Y)$ and $S(Y)$, respectively.
(iii) The GUM does not yield a pdf for the value of the measurand because the inputs are only expected values, variances and correlation coefficients. Thus it does not yield the coverage probability for uncertainty expressed as an interval. Some metrologists regard it as a limitation of the GUM. We do not fully concur with this view because uncertainty expressed as an interval with a stated coverage probability indicates exactitude which may not be warranted by the extent of factual information available concerning the input quantities.

These limitations of the GUM motivated the BIPM JCGM to develop the draft Supplement 1 to the GUM (the draft GUM-S1) on propagation of pdfs by Monte Carlo simulation [27].

5. The draft Supplement 1 to the GUM

In this section, we review the draft GUM-S1 and discuss its merits and limitations.

5.1. Measurement equation in the draft GUM-S1

In the draft GUM-S1, the measurement equation is a function $Y = f(X_1, \ldots, X_N)$ that represents the after-measurement-method of determining a state-of-knowledge pdf for the value Y of the measurand from a completely specified joint pdf for various input variables $X_1, \ldots, X_N$. The draft GUM-S1 requires the user to specify a state-of-knowledge pdf for each input variable $X_1, \ldots, X_N$ as well as to specify the joint pdf for $X_1, \ldots, X_N$. Therefore, the draft GUM-S1 requires more information concerning the input quantities than the GUM.

5.2. Propagation of pdfs by Monte Carlo simulation

The draft GUM-S1 [2] propagates the pdfs assigned to the input variables $X_1, \ldots, X_N$ to determine a pdf for the value Y of the measurand through a Monte Carlo simulation of the measurement equation $Y = f(X_1, \ldots, X_N)$. A result y and its associated standard uncertainty $u(y)$ for the value of the measurand are then determined from the simulated pdf for Y. The basic algorithm is as follows.

(1) Specify the measurement equation $Y = f(X_1, \ldots, X_N)$.
(2) Specify a joint pdf for $X_1, \ldots, X_N$. The draft GUM-S1 addresses two situations: (i) the input variables $X_1, \ldots, X_N$ are all mutually independent in which case the joint pdf is the product of the individual pdfs and (ii) the joint pdf for $X_1, \ldots, X_N$ is multivariate normal.
In principle, the draft GUM-S1 is applicable for any joint pdf for $X_1, \ldots, X_N$ from which random samples can be numerically generated and the corresponding values for Y calculated from the measurement equation. The draft GUM-S1 mentions multivariate normal distribution because numerical methods for generating random samples from this distribution are known.
(3) Generate M random samples $x_1^{(r)}, \ldots, x_N^{(r)}$, for $r = 1, \ldots, M$, from the joint pdf for $X_1, \ldots, X_N$ by Monte Carlo simulation. The number M may be set in advance or determined adaptively in real-time simulation. The draft GUM-S1 suggests that $M = 10^6$ is likely to be adequate.
(4) Calculate the M simulated values $y^{(r)} = f(x_1^{(r)}, \ldots, x_N^{(r)})$ for Y using the measurement equation.
(5) Calculate the result y and standard uncertainty $u(y)$ as the arithmetic mean and the experimental standard deviation of the M simulated values $y^{(1)}, \ldots, y^{(M)}$ for Y.
(6) Calculate a coverage interval $[y_{\text{low}}, y_{\text{high}}]$ for Y by determining the limits y_{low} and y_{high} such that the interval $[y_{\text{low}}, y_{\text{high}}]$ encompasses the desired fraction p of the simulated distribution for Y. The fraction p is the coverage probability of the interval. When the pdf for Y is asymmetric, the preferred interval $[y_{\text{low}}, y_{\text{high}}]$ may not be symmetric with respect to the result y. A common value for p is 95%.

The draft GUM-S1 discusses two forms of the interval $[y_{\text{low}}, y_{\text{high}}]$. The first form is the interval $[y_{(0.025)}, y_{(0.975)}]$, where $y_{(0.025)}$ and $y_{(0.975)}$ are the 0.025th quantile (2.5% percentile) and 0.975th quantile (97.5% percentile) of the simulated distribution for Y. The interval $[y_{(0.025)}, y_{(0.975)}]$ excludes equal probability 0.025 (2.5%) on each side and it is therefore a probabilistically symmetric interval. The second form is the shortest width interval $[y_{\text{low}}, y_{\text{high}}]$ having the coverage probability $p = 95\%$. The draft GUM-S1 seems to favour the shortest width interval $[y_{\text{low}}, y_{\text{high}}]$.

5.3. Type A and Type B probability distributions

The GUM applies Type A and Type B classification to the methods for evaluating the expected values, standard deviations and correlation coefficients of state-of-knowledge pdfs for the input variables of a measurement equation [1, 18]. A natural extension for the draft GUM-S1 is to apply Type A and Type B classification to the methods for specifying the input state-of-knowledge pdfs. Thus a Type A pdf is specified on the basis of statistical analysis of the current measurement data and a Type B pdf is specified by other means. Different approaches are needed for specifying the two types of pdfs. Generally, a Type A pdf is a Bayesian posterior probability distribution based on the current measurement data. A Type B pdf is specified by other means such as scientific judgment. The probabilistic interpretation of a Type B pdf is like that of a prior pdf used in Bayesian statistics. The draft GUM-S1 [1, section 5.11.3] states that a classification of state-of-knowledge pdfs into Type A and Type B is not needed. We beg to differ. Even though the draft GUM-S1 treats Type A and Type B pdfs in exactly the same way, this classification is useful for interpreting the output of numerical simulation and in discussing the relationship between the draft GUM-S1 and Bayesian statistics [25].

5.4. Merits of the draft GUM-S1

(i) When the measurement equation is non-linear, the draft GUM-S1 enables determination of the result y uncertainty $u(y)$ and coverage interval $[y_{\text{low}}, y_{\text{high}}]$ with a stated coverage probability for those joint input pdfs for which numerical simulation can be carried out.

(ii) When the measurement equation is non-linear, the draft GUM-S1 can be used to assess errors in the result y and standard uncertainty $u(y)$ determined from (13) and (14) for those joint input pdfs for which numerical simulation can be carried out.

(iii) The draft GUM-S1 can be used to assess the coverage probabilities of expanded uncertainty intervals $[y \pm k\,u(y)]$ determined from the GUM for those joint input pdfs for which numerical simulation can be carried out.

The draft GUM-S1 is based on Monte Carlo simulation. A Monte Carlo simulation is useful in evaluating uncertainty in measurement for multivariate measurands as well as in least-square computations [28–30].

5.5. Limitations of the draft GUM-S1

(i) A simulated pdf for the value Y of the measurand may give an impression of being factual and exact. However, it is factual only to the extent that the joint pdf assigned to the input variables is factual. All determinations of the pdfs for the input variables are based on assumptions and models which could be faulty. Suppose a scaled and shifted t-distribution is assigned as a Bayesian posterior distribution to an input quantity determined from a series of measurements (Type A). This assignment requires that the measurement procedure be in a state of statistical control with a fixed normal distribution. As noted earlier, this ideal state is impractical or difficult to achieve and maintain. To the extent that the actual measurement procedure deviates from the assumed ideal state, the assigned t-distribution and the simulated pdf for Y may not be factual. Likewise, a Type B pdf must be professionally determined and based on factual information; otherwise it may not be reliable. It would be useful to include in the draft GUM-S1 a discussion of the challenges in assigning a joint pdf to the input variables.

(ii) Sometimes it is possible to specify correlation coefficients between state-of-knowledge probability distributions for correlated input variables; however, often it is not possible to specify the joint pdf for the input variables or the joint pdf may not be in a form that is easy to numerically simulate. Then the GUM-S1 cannot be used directly. This issue is discussed in [28].

(iii) Uncertainty budget is a vital output from the calculation of uncertainty and it is a critical aid in assessing the combined standard uncertainty from the practical viewpoint [22,23]. An uncertainty budget requires sensitivity coefficients and individual components of uncertainty. The draft GUM-S1 [2, annex B] discusses how sensitivity coefficients may be numerically determined. However, an uncertainty budget is not a direct output of propagating pdfs by numerical simulation.

(iv) A simulated pdf for the value Y of the measurand determined from the draft GUM-S1 is not easily transferable. In the GUM, the inputs and the outputs are similar entities: expected values, standard deviations and correlation coefficients. In the draft GUM-S1, the pdf for each input variable is specified by a few parameters [2, section 6]. The output pdf for Y is a simulated distribution specified by M values where M may be to the order of 10^6. Thus the input and output are dissimilar entities. The simulated values for Y cannot be easily archived and transmitted for subsequent use as input for another uncertainty calculation on a different occasion by another person. If only the result, standard uncertainty and coverage interval are reported then the simulated pdf cannot be reconstructed from the reported values. However, a result and standard uncertainty obtained from the draft GUM-S1 may be used as inputs for another evaluation of uncertainty according to the GUM.

(v) As discussed in the draft GUM-S1, Monte Carlo simulation involves many numerical issues. In particular, if the computational algorithm is not developed carefully, numerical errors can accumulate. Thus professionally developed, good and easy to use computational software which avoid accumulation of numerical errors are needed. The GUM also requires computational software; however, the numerical issues in the draft GUM-S1 are more complex and challenging.

We regard the draft GUM-S1 as a useful companion to the GUM.

5.6. Relationship between the draft GUM-S1 and Bayesian statistics

A topic of current interest among leading researchers in metrology is the relationship between the draft GUM-S1 and Bayesian statistics. Reference [25] addressed this topic in the context of linear calibration. We regard the following books as authoritative references on Bayesian statistics: [31–36]. Bayesian statistics provides inference about unknown parameters on the basis of available statistical data and prior information. The link between the data and the unknown parameters is provided by a likelihood function. The likelihood function is the sampling pdf of the data regarded as a function of the parameters. Thus Bayesian statistics is a method for statistical (Type A) evaluation of data. The draft GUM-S1 goes beyond Bayesian statistics to combine statistical evaluations (Type A) and non-statistical evaluations (Type B) through a Monte Carlo simulation. In this sense, the draft GUM-S1 is an extension of Bayesian statistics.

6. Summary

The ideas of error and uncertainty were mixed up until the GUM clarified their meanings. Error analysis did not yield a combined expression of uncertainty that is needed in tabulating results of measurement with uncertainties. A report from error analysis provides information on imprecision and bound on bias. A user of the report is expected to use this information to determine his/her uncertainty in light of the intended use. There is arbitrariness in the classification of the sources of error as random and systematic, yet in error analysis the two components of error are treated differently. These and other limitations of the error analysis were hindrance to communication of scientific and technical measurements. Therefore the world's leading authorities in metrology discussed the issues and developed the GUM to supplant error analysis for quantifying uncertainty in measurement. A concept of enduring importance from error analysis is that the measurement procedure must be in a state of statistical control otherwise the measurement data cannot be treated statistically. Until the technology of measurement improves to such an extent that the random errors become negligible, the need for statistical control of the measurement procedure will remain.

While error analysis is based on the hypothetical concepts of true value and error in the realm of the state of nature, the GUM deals with quantifiable result of measurement and its associated uncertainty in the realm of the state of knowledge about nature. The GUM introduced the concept of measurement equation to quantify a result of measurement and its associated combined standard uncertainty. Before publication of the GUM, this concept did not exist in the literature on metrology or on statistical methods. The GUM recognizes the importance of scientific judgment in evaluating components of uncertainty and explicitly legitimized its use. The GUM declared that no distinction be made between statistical (Type A) and other (Type B) methods of evaluation. Before publication of the GUM there was no generally accepted approach to quantify a realistic expression of uncertainty from systematic effects. The GUM promulgated the principle that a result of measurement should be corrected for all recognized significant systematic effects and that every effort should be made to identify such effects. A correction is always uncertain. So the uncertainty associated with each correction should be quantified and included in the combined standard uncertainty. The GUM regards the variations arising from random effects and those associated with the corrections for systematic effects as components of uncertainty. The GUM recommends a logically consistent way of combining all components of uncertainty.

The GUM propagates the estimates, standard uncertainties and correlation coefficients for various input quantities through a linear approximation of the measurement equation to determine an estimate and standard uncertainty for the value of the measurand. Thus, the GUM is applicable as an approximation in those situations where reasonable estimates for the expected values, standard deviations and correlation coefficients are available. The GUM does not require a more extensive knowledge of the complete state-of-knowledge pdfs for the input variables. A limitation of the GUM is that when the measurement equation is non-linear the standard uncertainty for the value of the measurand determined by using the GUM may be a poor approximation. This and other limitations of the GUM motivated the BIPM JCGM to develop draft supplement 1 to the GUM (draft GUM-S1).

The draft GUM-S1 propagates the pdfs assigned to the input quantities to determine a pdf for the value Y of the measurand through a Monte Carlo simulation of the measurement equation. The draft GUM-S1 enables determination of the result y, uncertainty $u(y)$ and coverage interval $[y_{\text{low}}, y_{\text{high}}]$ from a linear or a non-linear measurement equation for those joint pdfs for the input variables for which numerical simulation can be carried out. A limitation of the draft GUM-S1 is that it requires knowledge of the joint distribution of correlated input variables which in complex situations may not be known. Also, it does not yield an uncertainty budget directly. Further, a simulated pdf for the value of the measurand cannot be easily archived and transmitted for subsequent use. Despite its limitations, the draft GUM-S1 is a useful companion to the GUM.

We hope this discussion will lead to more effective use of the GUM and its companion the draft GUM-S1. Also, we hope this discussion will stimulate investigations to improve these approaches to quantify uncertainty in measurement.

Acknowledgment

The authors thank the anonymous referee, Tyler Estler and Bert Rust for their comments on an earlier draft of this paper.

Appendix

NBS COMMUNICATIONS MANUAL FOR SCIENTIFIC, TECHNICAL, AND PUBLIC INFORMATION

Edited by
Carol W. Solomon and Randall D. Bograd
Writing Consultants
and
W. **Reeves Tilley**
National Bureau of Standards

***Chapter 15 of the NBS Administrative Manual**

This Manual is based on material prepared by an Ad Hoc Committee consisting of Edward Brady, Robert Parker, Carl Muehlhause, Sam Chappell, Reeves Tilley, Robert Blunt, Paul Campbell, Dick Franzen, Gordon Day, and Ralph Desch. It supersedes the NBS Publications and Reports Manual of June 1969 (and subsequent changes), and Chapter 15, dated 1968 through 1976, of the NBS Administrative Manual.

For Official Distribution / November 1980

2-30 **POSTSCRIPT**

Over the intervening years since the publication of Eisenhart's and Ku's articles, it has become apparent that a few additional comments may be useful. It is equally apparent that a complete revision is neither necessary nor desirable inasmuch as the major thrust and content of the articles remain as valid and as appropriate as when first written. For this reason, these comments are made as a postscript.

Uncertainty Assessments Must Be Complete

The uncertainty of a reported value is meant to be a credible estimate of the likely limits to its actual *error, i.e.,* the magnitude and sign of its deviation from the truth. As such, uncertainty statements must be based on as nearly complete an assessment as possible. This assessment process must consider every conceivable source of inaccuracy in the result.

A measurement process generally consists of a very complicated sequence of many individual unit operations or steps. Virtually every step in this sequence introduces a conceivable source of inaccuracy whose magnitude must be assessed. These sources include:

- Inherent stochastic variability of the measurement process;
- Uncertainties in standards and calibrated apparatus;
- Effects of environmental factors, such as variations in temperature, humidity, atmospheric pressure, and power supply voltage;
- Time-dependent instabilities due to gradual and subtle changes in standards or apparatus;
- Inability to realize physical model because of instrument limitations;
- Methodology procedural errors, such as incorrect logic, or misunderstanding what one is or should be doing;
- Uncertainties arising from interferences, impurities, inhomogeneity, inadequate resolution, incomplete discrimination, etc.;
- Metrologist errors, such as misreading of an instrument;
- Malfunctioning or damaged apparatus;
- Laboratory practice including handling techniques, cleanliness, etc.; and
- Computational uncertainties as well as errors in transcription of data, and other calculational or arithmetical mistakes.

This list should not be interpreted as exhaustive, but rather as illustrative of the most common generic sources of inaccuracy that may be present.

The various sources of inaccuracy are generally classified into sources of *imprecision* (random components) and sources *of bias* (fixed offsets). To which category a particular source should be properly assigned is often difficult and troublesome. In part, this is because many experimental procedures or individual steps in the overall measurement process embody both systematic and stochastic (random) elements. (For an alternative discussion that questions the need for a clear cut distinction between random and systematic components of uncertainty, see [7].) One practical approach is to classify the sources of inaccuracy according to how the uncertainty is estimated. In this way, sources of imprecision are considered to be those components which *can be* and *are* estimated by a statistical analysis of replicate determinations. For completeness, the *systematic uncertainty components* can be considered to be the residual set of conceivable sources of inaccuracy that are biased and not subject to random variability, and those that may be due to random causes but *cannot be* or *are not* assessed by statistical methods. The systematic category includes sources of inaccuracy other than biases in order to obtain a complete accounting of all sources of inaccuracy in the measurement process. Hence, it is meaningful to report a random uncertainty contribution, only if one has a computed statistic for the magnitude of its imprecision or random variation. Many sources of inaccuracy may exist consisting of several components from both the random and systematic categories and can be assessed only after consideration of the more fundamental processes involved. The uncertainty in the calibration of an instrument with a standard reference material, for example, would have not only components from the uncertainty in the standard itself, but also uncertainty components arising from the use of the standard in performing the calibration.

Assessment of Imprecision (Random Uncertainties)

Although the treatment and expressions of reporting the imprecision of measurement results were adequately covered in the original article, a number of points are of sufficient importance to deserve reemphasis.

The only way to assess realistically the overall imprecision is to make direct—or preferably, when possible, indirect—replicate determinations [1] and calculate an appropriate statistic such as the standard error of the mean. It is extremely important to be definite on what constitutes a "replicate determination" because the extent to which conditions are allowed to vary freely over successive "repetitions" of the measurement process determines the scope of the statistical inferences that may be drawn from measurements obtained [2, sec. 4.1]. When measurements of a particular quantity made on a single occasion exhibit closer mutual agreement than measurements made on different occasions so that differences between occasions are indicated, the value of the computed standard error of the mean of all the measurements obtained by lumping all of the measurements together will underestimate the actual standard error of the mean. A more realistic value is given by taking the arithmetic means of the measurements obtained on the respective occasions as the *replicate determinations* and calculating the standard error of their mean in the usual way [3, sec. 3.5].

In many situations, it may not be possible or feasible because of time and cost constraints to perform a sufficient number of completely independent determinations of the measurement result. For results derived from several component quantities, the individual imprecision estimates must be propagated to obtain the imprecision of the final result. It must be emphasized, however, that these estimates of imprecision should not be based exclusively on the information derived from just the present measurements. Presently derived information should be added to the information accumulated in the past on the imprecision of the measurement process. In this way, more realistic and reliable canonical values of the imprecision statistics may be established over time. Ideally, every major step or component of the measurement process should be independently assessed. This would include not only the variability inherent in the particular measurement of concern, but also the imprecision arising from corrections, calibration factors, and any other quantities that make up the final result.

Assessment of Systematic Uncertainties

Although a general guideline for the approach to the assessment of systematic uncertainties can be formulated, there are, unfortunately, no rules to objectively assign a magnitude to them. For the most part, it is a subjective process. Their magnitudes should preferably be based on experimental verification, but may have to rely on the judgment and experience of the metrologist. In general, each systematic uncertainty contribution is considered as a quasi-absolute upper bound, overall or maximum limit on its inaccuracy. Its magnitude is typically estimated in terms of an interval from plus to minus δ about the mean of the measurement result. By what method then should the magnitude of these maximum limits be assigned? It may be based on comparison to a standard, on experiments designed for the purpose [4], or on verification with two or more independent and reliable measurement methods. Additionally, the limits may be based on judgment, based on experience, based on intuition, or based on other measurements and data. Or the limits may include combinations of some or all of the above factors. Whenever possible, they should be empirically derived or verified. The reliability of the estimate of the systematic uncertainty will largely depend on the resourcefulness and ingenuity of the metrologist.

The Need for an Overall Uncertainty Statement

Without deprecating the perils of shorthand expressions, there is often a need for an overall uncertainty statement which combines the imprecision and systematic uncertainty components. Arguments that it is incorrect from a theoretical point of view to combine the individual components in any fashion are not always practical. First, an approach which retains all details is not amenable for large compilations of results from numerous sources. And second, this approach shifts the burden of evaluating the uncertainties to users. Many users need a single uncertainty value resulting from the combination of all sources of inaccuracy. These users believe, and rightly so, that this overall estimate of inaccuracy can be most appropriately made by the person responsible for the measurement result. It must be emphasized, however, that there is no one clearly superior appropriate method for reporting an overall uncertainty, and that the choice of method is somewhat arbitrary. Several methods are commonly employed [5,6].

One method is to add linearly all components of the systematic uncertainty and linearly add the total to the imprecision estimate. Since the individual systematic uncertainties (δ_j) are considered to be maximum limits, it logically should be added to an imprecision estimate at a similar confidence level. That is, for example, the overall uncertainty u may be given by

$$u = [t_v(\alpha)]s + \sum_{j=1}^{q} \delta_j$$

where s is the computed standard error based on v degrees of freedom, $t_v(\alpha)$ *is* the Student-*t* value corresponding to a two-tail significance level of α=0.05, 0.01, or 0.001 (depending on the practice in the measurement field concerned), and δ_j is the magnitude of the estimated systematic uncertainty for each of the identified q systematic uncertainty components. This approach probably overestimates the inaccuracy, but can be considered as an estimate of the maximum possible limits. For example, if someone estimated that five contributions of about equal magnitude made up the total systematic error, that person would have to be very unlucky if all five were plus, or all five were minus. Yet, if there was one dominant contributor, it might be a very valid approximation.

Two other approaches have also been widely used. These methods add in quadrature all of the systematic uncertainty components, and either add the resulting quantity *linearly* to the standard error estimate,

$$s + \sqrt{\sum_{j=1}^{q} \delta_j^2} \ ,$$

or add it *in quadrature* to the standard error estimate,

$$\sqrt{s^2 + q\sum_{j=1} \delta_j^2}$$

These are frequently considered (erroneously) to correspond to a confidence level with P=68%.

In another method, often termed the PTB approach [6], the component systematic uncertainties are assumed to be independent and distributed such that all values within the estimated limits are equiprobable (rectangular or uniform distribution) [8]. With these assumptions, the rectangular systematic uncertainty distributions can be convoluted to obtain a combined probability distribution for which the variance may be computed. This may then be combined in quadrature with that for the random uncertainty. In its simplest form, the uncertainty components are combined to form an overall uncertainty by

$$u = k\sqrt{s^2 + (1/3)\sum_{j=1}^{q} \delta_j^2} \ ,$$

where k is customarily taken as 2 or 3. The above simple form is not appropriate when one of the component δ_j's is much larger than the others; in such a case it will be more informative to keep that component separate from the others and add it linearly.

Exhibit 2-E *(cont.)*

2-32

A Concluding Thought

If there is one fundamental proposition for the expression of uncertainties, it is

The information content of the statement of uncertainty determines, to a large extent, the worth of the final result.

This information content can be maximized by following a few simple principles:

BE EXPLICIT

PROVIDE DETAILS

DON'T OVERSIMPLIFY

When an overall uncertainty is reported, one should explicitly state how the separate components were combined. In addition, for results of primary importance, a detailed discussion and complete specification of all of the separate uncertainty components is still required. In this way, some users will benefit from having the metrologist's estimate of the overall uncertainty, while more sophisticated users will still have access to all of the information necessary for them to evaluate, combine, or use the uncertainties as they see fit.

REFERENCES AND NOTES

[1] Youden, W. J. Statistical aspects of analytical determinations. The Analyst 77: 874-878; 1952, Dec: Reprinted in Journal of Quality Technology 4: 45-49; 1972, Jan: Youden, W. J., Connor, W. S., Making one measurement do the work of two; Chemical and Engineering Progress 49: 549-552; 1953, Oct. Reprinted in Journal of Quality Technology 4: 25-28; 1972, Jan.

[2] Eisenhart, Churchill. Realistic evaluation of the precision and accuracy of instrument calibration systems. J. Res. Nat. But. Stand. (U.S.) 67C (2): 161-187; 1963; Reprinted as paper 1.2 in NBS Special Publication 300-1.

[3] Eisenhart, Churchill. Contribution to panel discussion of adjustments of the fundamental physical constants. Langenberg, D. N., Taylor, B. N., eds. Precision Measurement and Fundamental Constants, Nat. Bur. Stand. (U.S.) Spec. Publ. 343: 509-518; 1971.

[4] Youden's burette experiment, Journal of Quality Technology 4: 20-23 1972, Jan. Youden, W. J., Systematic errors in physical constants. Physics Today, 14, 32-34, 36, 38, 40, 43; 1961, Sept. Reprinted as paper 1.4 in NBS Special Publication 300-1. Youden, W. J., Enduring values, Technometrics 14: 1-11; 1972, Feb.

[5] Campion, P. J.; Burns, J. E.; Williams, A. A code of practice for the detailed statement of accuracy. National Physical Laboratory. London: Her Majesty's Stationery Office. 1953. II-57.

[6] Wagner, Siegfried R. On the quantitative characterization of the uncertainty of experimental results in metrology: PTB-Mitteilungen 89: 83-89; 1979, Feb.

[7] Müller, Jörg G. Some second thoughts on error statements. Nuclear Instruments and Methods 163, 241-251; 1979.

[8] A numerical comparison of uncertainty limits resulting from these assumptions with those implied by several alternative distributional assumptions is provided by table 1 on page 184 of [2], and discussed on the same and following page.

Churchill Eisenhart
Ronald Collé
July 1980

Exhibit 2-E *(cont.)*

References

[1] 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd edn (Geneva: International Organization for Standardization)
[2] *Draft GUM Supplement 1: Propagation of Distributions using Monte Carlo Method* (BIPM Joint Committee on Guides in Metrology JCGM)
[3] *The NIST Reference on Constants, Units, and Uncertainty* http://physics.nist.gov/cuu/index.html
[4] Müller J W 1984 The assignment of uncertainties to the results of experimental measurements in *Precision Measurement and Fundamental Constants II* ed B N Taylor and W D Phillips (Gaithersburg, MD: National Bureau of Standards (US)) Spec. Pub. 617
[5] Eisenhart C 1963 Realistic evaluation of the precision and accuracy of instrument calibration systems *J. Res. Natl Bur. Stand.* C **67** 161–87
[6] Eisenhart C 1968 Expression of the uncertainties of final results *Science* **160** 1201–4
[7] Ku H H 1968 Expression of imprecision, systematic error, and uncertainty associated with a reported value *Meas. Data* **2** 72–7
[8] Ku H H 1966 Notes on the use of propagation of error formulas *J. Res. Natl Bur. Stand.—C Eng. Instrum.* **70** 263–73
[9] Bevington P and Robinson D K 1992 *Data Reduction and Error Analysis for the Physical Sciences* 2nd edn (New York: McGraw Hill)
[10] Taylor J R 1997 *An Introduction to Error Analysis* 2nd edn (Sausalito, CA: University Science Books)
[11] Youden W J 1972 *J. Qual. Technol.* **4** (American Society for Quality Control) (a special issue in honor of Jack Youden)
[12] Youden W J 1961 Systematic errors in physical constants *Phys. Today* **14** 32–42
[13] Natrella M G 2005 *Experimental Statistics* (New York: Dover)
[14] Müller J W 1979 Some second thoughts on error statements *Nucl. Instrum. Methods* **163** 241–51
[15] Eisenhart C and Collé R 1980 Postscript to 'Expressions of imprecision, systematic error, and uncertainty associated with a reported value' *NBS Communications Manual for Scientific, Technical, and Public Information* ed Solomon *et al* (Gaithersburg, MD: US Department of Commerce, National Institute of Standards and Technology) Chapter 2 Exhibit 2-E
[16] Wagner S R 1979 On the quantitative characterization of the uncertainty of experimental results in metrology *PTB Mitt.* **89** 83–9
[17] Taylor B N and Kuyatt C E 1994 Guidelines for evaluating and expressing the uncertainty of NIST measurement results *NIST Technical Note* 1297 (Gaithersburg, MD: US Department of Commerce, National Institute of Standards and Technology)
[18] Kacker R N and Jones A T 2003 On use of Bayesian statistics to make the *Guide to the Expression of Uncertainty in Measurement* consistent *Metrologia* **40** 235–48
[19] Sommer K D and Siebert B R L 2006 Systematic approach to the modelling of measurements for uncertainty evaluation *Metrologia* **43** S200–10
[20] Eppeldauer G P, Sauter G and Gardner J L 2006 Uncertainties of spectral responsivity measurements *Proc. 2nd CIE Expert Symp. on Measurement Uncertainty* (Vienna: CIE Central Bureau) pp 133–8
[21] Gardner J L 2003 Uncertainties in interpolated spectral data *J. Res. Natl Inst. Stand. Technol.* **108** 69–8
[22] 1999 *Expression of the Uncertainty of Measurement in Calibration* EA-4/02 (European Co-operation for Accreditation, www.european-accreditation.org)
[23] 2000 *Quantifying Uncertainty in Analytical Measurement* 2nd edn (Eurachem/Citac, http://www.measurementuncertainty.org)
[24] Kessel R, Kacker, R N and Berglund M 2006 Coefficient of contribution to the combined standard uncertainty *Metrologia* **43** S189–95
[25] Kacker R N, Toman B and Huang D 2006 Comparison of GUM, draft GUM Supplement 1 and Bayesian statistics using simple linear calibration *Metrologia* **43** S167–77
[26] Kacker R N 2006 Bayesian alternative to the GUM's use of the Welch–Satterthwaite formula *Metrologia* **43** 1–11
[27] Bich W, Cox M G and Harris P M 2006 Evolution of the *Guide to the Expression of Uncertainty in Measurement Metrologia* **43** S161–6
[28] Cox M G and Siebert B R L 2006 The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty *Metrologia* **43** S178–88
[29] Hammersley J M and Handscomb D C 1965 *Monte Carlo Methods* (London: Chapman and Hall)
[30] Gardner J L 2006 Uncertainties in source distribution temperature and correlated colour temperature *Metrologia* **43** 403–8
[31] Box G E P and Tiao G C 1973 *Bayesian Inference in Statistical Analysis* (Reading, MA: Addison-Wesley)
[32] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* 2nd edn (Berlin: Springer)
[33] Bernardo J M and Smith A F M 1994 *Bayesian Theory* (New York: Wiley)
[34] Lee P M 1997 *Bayesian Statistics An Introduction* 2nd edn (Oxford: Oxford University Press)
[35] Press S J 2003 *Subjective and Objective Bayesian Statistics* 2nd edn (New York: Wiley)
[36] Gelman A, Carlin J B, Stern H S and Rubin D B 2004 *Bayesian Data Analysis* 2nd edn (London: Chapman and Hall)