Metrologia

Simple informative prior distributions for Type A uncertainty evaluation in metrology

To cite this article: Anthony O'Hagan and Maurice Cox 2023 *Metrologia* **60** 025003

OPEN ACCESS
IOP Publishing | Bureau International des Poids et Mesures Metrologia
Metrologia **60** (2023) 025003 (17pp) https://doi.org/10.1088/1681-7575/acb93d

Simple informative prior distributions for Type A uncertainty evaluation in metrology

Anthony O'Hagan[1] **and Maurice Cox**[2,*]

[1] University of Sheffield, Sheffield, United Kingdom
[2] National Physical Laboratory, Teddington, United Kingdom

E-mail: maurice.cox@npl.co.uk

Received 27 August 2022, revised 18 January 2023
Accepted for publication 6 February 2023
Published 22 February 2023

Abstract

The result of a measurement, including the expression of uncertainty in the measurement, should represent a carefully considered opinion based on the metrologist's experience and expertise, as well as on the data and other information sources. This is the position of the Guide to the expression of uncertainty in measurement (GUM), where the requirement for such judgment is clear in the case of Type B (non-statistical) evaluation. However, when making Type A evaluations, involving statistical analysis of data, the GUM and related GUM documents implicitly consider the data to be the only relevant information. This perspective is unfortunate, and arguably unscientific, when, as is frequently the case, the metrologist could bring other relevant information to bear. Bayesian statistical methods allow the use of prior information in addition to the data in Type A evaluation and have been advocated by several authors. However, prior information is in principle subjective and, as in other fields, there is some resistance in the metrology community at large to embrace Bayesian methods using meaningful, subjective prior probability distributions. We address our paper to metrologists in measurement and calibration laboratories whose workload is such that new techniques will only be adopted if they have proven advantages and are straightforward to apply routinely. We present two prior distributions for use in the most basic of all Type A evaluations, where the data comprise a sample of indications assumed to be normally distributed. These distributions represent prior information about the observation error variance in a simple form that is readily justified in practice. We show the gains to be achieved by using these prior distributions, both in the single Type A evaluation and in a more complex measurement model, and present simple guidance for verifying their validity.

Keywords: Type A uncertainty evaluation, measurement uncertainty, GUM, Bayesian methods, prior information, characteristic uncertainty

(Some figures may appear in colour only in the online journal)

* Author to whom any correspondence should be addressed.

1. Introduction

1.1. The guide to the expression of uncertainty in measurement

Two components, the estimate of a quantity and the associated measurement uncertainty, together constitute a common way of reporting a measurement result [1].

The Joint Committee for Guides in Metrology (JCGM) maintains and promotes the Guide to the expression of uncertainty in measurement (GUM) [2]. The GUM (JCGM 100) has for many years been the authoritative guide concerned with the evaluation of measurement uncertainty. The GUM is one part of a suite of documents including [3–6] that use the concept of *standard uncertainty*, defined [2, clause 2.3.1] as the 'uncertainty of the result of a measurement expressed as a standard deviation'.

The GUM treats measurement as in general involving a *measurement model* relating the measurand Y (taken as a univariate quantity here) to input quantities X_i:

$$Y = f(X_1, \ldots, X_N).$$

Knowledge of Y is determined given f and estimates x_i of the X_i, associated standard uncertainties $u(x_i)$ and possibly covariances between the X_i. The process of determining the estimate and standard uncertainty of a model input is itself a measurement, and in the context of such a measurement we will refer to the input quantity as the measurand.

The GUM gives guidance on providing estimates x_i and on Type A and Type B evaluation of the X_i. A Type A evaluation of X_i uses statistical methods such as taking the mean of a set of indications obtained independently under the same measurement conditions, and using the standard error of the mean as $u(x_i)$. A Type B evaluation of X_i uses a knowledge-based probability distribution for an input quantity, taking the standard deviation of the distribution as $u(x_i)$.

The evaluation of X_i occurs prior to employing the measurement model in which X_i is to be used as an input to a measurement of Y. Now X_i is no longer the subject of the measurement and, in the context of that measurement, Y is the measurand and X_i is just an input.

Degrees of freedom are assigned in the GUM to inputs as part of Type A and Type B evaluations. A probability distribution is characterized by the measurement result as a normal or a shifted and scaled Student's t distribution from which a confidence interval for the measurand is obtained.

Confidence intervals are expressed in terms of expanded uncertainty. So, a typical measurement result will involve an estimate, a standard uncertainty and an expanded uncertainty for 95 % coverage.

Although the GUM has huge influence in metrology, with measurement uncertainty being routinely evaluated according to the above procedure, several key components of that procedure have been challenged. We subsequently set out alternative elements that will be adopted in this article.

1.2. Two statistical paradigms

A long-running controversy in metrology concerns the underlying statistical methodology employed for the expression of uncertainty in measurement. The two principal paradigms in statistics, termed 'frequentist' and 'Bayesian', express uncertainty in different ways and using different formal definitions of probability.

- Frequentist methods are based on the frequency definition of probability, where the probability of an event is defined to be the frequency with which that event occurs in the long run, over many repetitions. The Type A procedures given in the GUM are based on frequentist statistical theory, and accordingly the resulting standard uncertainties quantify how variable the estimate of a measurand will be over many repetitions of the measurement process.
- Bayesian methods employ a subjective definition of probability, whereby the probability of an event is a subjective judgment representing a person's rational degree of belief that it will occur. Type B evaluation in the GUM is a subjective judgment and the resulting standard uncertainty quantifies the metrologist's knowledge of the measurand.

When Type A and Type B evaluations are combined, the GUM mixes frequentist and Bayesian concepts and has received considerable criticism for doing so (references include [7–9]). We believe that the only logical, coherent solution is to adopt the Bayesian paradigm consistently, including making Type A evaluations using Bayesian methods. Such a practice accords with the suggestion in the GUM [2, clause E.3.5] that disparate standard uncertainties can be combined because ultimately all expressions of uncertainty must be the metrologist's judgment and opinion, and with the use of Bayesian methods in JCGM 101.

1.3. Bayesian methods

From the Bayesian perspective, uncertainty in any quantity is expressed using probabilities, and a probability distribution represents a complete description of that uncertainty. A Bayesian Type A evaluation for a quantity X will therefore result in a probability distribution for X. It should represent the metrologist's considered judgments about X based on all available information. In Bayesian analysis, a distinction is made between the data, typically comprising the sample of experimental determinations of X for a Type A evaluation, and the prior information, comprising all other knowledge the metrologist may have, including experience with the measurement procedure and with quantities such as X in the past. The information in the data is represented through the same statistical model as would be used in frequentist Type A evaluation, but the Bayesian analysis also recognizes the prior information represented as a prior distribution for X. The two sources of knowledge are synthesized in a natural way using Bayes' theorem. The result is a probability distribution for X,

the *posterior* distribution, representing the sum of the metrologist's knowledge about X.

An important potential advantage of Bayesian methods in metrology is the ability to make use of more information. The addition of prior knowledge will typically result in less uncertainty regarding X than would have been obtained through use of the data alone. It is worth noting that day-to-day measurements in practising laboratories often involve Type A evaluations with very few experimental determinations; sample sizes as low as three or four are commonplace. In this context, the addition of prior information can offer valuable improvements.

Where a measurand is expressed as a function of various input quantities through a measurement model, these quantities will all have probability distributions in a Bayesian analysis. An input that is subject to Type A evaluation will have a posterior distribution. One that is subject to Type B evaluation will have a distribution expressed directly as the metrologist's judgment. Distributions for the inputs to a measurement model imply a probability distribution for the measurand, which may for instance be computed using the Monte Carlo method of JCGM 101 [3]. Probabilities and probability distributions are always to be understood as representing the considered opinion and judgment of the metrologist.

Scientists whose training in statistics has been confined to the more common frequentist concepts and methods are not accustomed to applying a Bayesian treatment. Further, although Bayes' theorem is a standard statistical tool, it involves probability operations that are unfamiliar to many practitioners. Bayesian methods are therefore often regarded as more complex and more mathematical than frequentist methods, but this is an unfair perception. One purpose of this article is to demonstrate that Bayesian methods can be equally straightforward to apply in practice.

1.4. Prior information

The potential advantages of using Bayesian methods in metrology are two-fold. First, adopting the Bayesian framework provides a rigorous and legitimate way to combine Type A and Type B evaluations. Second, the incorporation of the metrologist's prior knowledge in Type A evaluation may strengthen the measurement and allow a more realistic uncertainty to be reported from a given sample of data.

Bayesian methods require the specification of a prior distribution that represents the metrologist's judgment about the likely values of a quantity before seeing the data that will be used to obtain the measurement result, based on background knowledge and experience. As such, it is necessarily subjective; different metrologists in the same context might express different prior judgments, although it is the prior knowledge and professional judgment of the metrologist who is responsible for the measurement result that matters. Bayesian methods are widely used in almost all areas where statistical analysis is employed, but they often face resistance because of the subjective (individual) nature of the prior distribution. In metrology, the use of a prior distribution may be viewed, we believe unfairly, as compromising or undermining the objectivity of the data.

To address concerns about subjectivity, some practitioners of Bayesian statistics use so-called noninformative prior (NIP) distributions that are supposed to be objective representations of prior ignorance. In metrology, the standard deviation of the prior distribution expresses the strength of prior information. The larger is the standard deviation, the weaker is the prior information. A NIP distribution should therefore have a standard deviation that is so large as to be effectively infinite. Using such a prior distribution, it is claimed, should gain the first benefits of Bayesian methods, namely a rigorous framework for combining Type A and Type B evaluations, without contaminating the data with the metrologist's subjective judgments.

Although some practitioners might consider this approach attractive, it is controversial for several reasons:

- Numerous formulations have been proposed for representing the notion of prior ignorance, and in any given situation they may give different 'noninformative' distributions that lead to quite different posterior distributions. There is no unique, objective distribution to represent a state of complete or near ignorance regarding a measurand;
- In practice, there is always some prior knowledge. For example, without some idea of likely values for a measurand it is not possible to devise or utilize a suitable measurement procedure, and there is always prior knowledge about the error characteristics of any measuring system;
- Claiming prior ignorance when there is in fact some prior information implies failing to use all available information regarding the measurand. That may be seen as unscientific, and even a derogation of duty.

Ultimately, the notion that subjectivity is unacceptable in science is widespread but demonstrably false. Subjective judgment is a feature of all scientific activity: examples include formulating hypotheses, building models, designing experiments, choosing how to analyse data and interpreting data. Good science involves judgments and opinions being formed carefully and rigorously, *scientifically*, and being open to challenge in the forum of scientific debate and peer review.

It is our opinion, therefore, that where genuine prior knowledge exists it should be acknowledged and used, in the form of an informative prior distribution, and not denied by substituting a 'noninformative' distribution. An informative prior distribution will often allow the reporting of a smaller measurement uncertainty and a correspondingly narrower 95 % coverage interval. However, these benefits depend on the prior distribution being realistic.

A 95 % interval for an unknown quantity should contain the true value of that quantity with probability 0.95. The practical meaning of this statement in metrology is that over a long period of time, when a metrologist constructs many 95 % intervals for many measurands, then approximately 95 % of those intervals should contain the true values of those measurands. This statement will be true when the intervals are constructed

as *confidence intervals*, using the frequentist approach to statistics, provided that the statistical model used to construct the intervals is valid. It is shown in appendix A that it is also true for Bayesian intervals, but only on the additional condition that the prior distributions for those measurands are realistic. That is, the true measurands should behave as if they have been sampled from their prior distributions. If, for instance, the true values of the measurands were mostly to lie in the upper tails of the metrologist's prior distributions, then those prior distributions would not be realistic. The metrologist's prior judgments would consistently underestimate the measurands, and we would not expect 95 % of the resulting intervals to contain the true measurand values.

1.5. Target audience

The use of Bayesian methods in metrology, including the use of informative prior distributions, has been advocated and adopted in many publications such as [10–13]. Nevertheless, applications have largely been confined to researchers, for instance, in national metrology institutes, rather than practitioners, and have made little impact on measurement practice in the community at large.

The target audience of this paper is metrologists in measurement and calibration laboratories whose workload (and financial model) is such that new techniques will, quite legitimately, only be adopted if they have proven advantages and are simple to apply routinely. The paper is directed at metrologists who wish to use their knowledge of their measurement to reduce in a straightforward way the uncertainty in their measurement results.

We focus on Type A evaluation from a sample of normally distributed indications, because for our target audience this is by far the most widely used technique for the expression of uncertainty.

Some metrologists might consider that adopting a Bayesian approach means they must discard the methodologies they have developed according to the GUM. They can be reassured that this is not the case: Type B evaluation is already Bayesian, and Type A evaluation in the GUM is in many cases equivalent to a Bayesian approach with a NIP. In particular, for the Type A evaluation from a normally distributed sample, the GUM [2] approach can be viewed as providing a Student's t distribution equivalent to the Bayesian posterior, while the approach of JCGM 101 [3] can be viewed as sampling from the Bayesian posterior, each employing the same choice of NIP for the measurand.

However, in a Bayesian approach we are not restricted to using a NIP. In implementing such an approach, we can choose priors that are appropriate for the situation, in particular, those encoding useful information about the observation error variance that the metrologist may know, for instance from published performance data of the measuring system or from repeatability of previous experiments. There is a class of priors that encode precisely this type of information but lead to exactly the same class of posterior distributions (Student's t distributions) with which the GUM already deals. Bayes gives this added flexibility at no computational overhead, with the advantage that in many circumstances the reported uncertainty can be smaller because of the better use of information. We emphasize that a Bayesian treatment is more flexible and so better able to encode information, leading to smaller uncertainties.

1.6. Desiderata for informative prior distributions in metrology

We have argued that prior information can and should be used in metrology to enhance the specific data, and indeed that this is one of the important benefits of adopting a Bayesian paradigm. However, we have also seen that (a) metrologists have understandable concerns about the use of subjective prior information, (b) the performance of coverage intervals may be poor if the prior distribution is not valid, and (c) Bayesian methods are generally seen to be complex and mathematically or computationally demanding. These issues must be addressed if informative prior distributions are to find widespread practical application in metrology.

The following list of desirable criteria has been formulated specifically with our target audience in mind. While they may be seen as overly restrictive for metrologists wishing to embrace Bayesian methods in research and complex measurement problems, they also relate to aspects of research reproducibility [14].

- *Justification.* The prior distribution should be based on judgments that are open to scrutiny and justified by reference to prior information and experience.
- *Simplicity.* The Bayesian procedure that derives the distribution of the measurand and summary information such as standard uncertainty [2] should be documented in a peer-reviewed source. It should be simple to use and to be replicated.
- *Benefit.* The advantages of incorporating prior information, for instance, in terms of the use of domain knowledge (in addition to the data) and a reduced measurement uncertainty, should be sufficient to offset any time or resource implications in adopting the Bayesian procedure.
- *Verification.* The consistency of the prior distribution and the data should be capable of verification.

1.7. Outline of the paper

Section 2 considers the most widely used case of Type A evaluation, in which the data comprise a sample of independent determinations with normally distributed observation errors. We consider two simple informative prior distributions that represent the kind of prior information that a metrologist will typically have, from previous experience and knowledge of the measurement procedure, regarding the magnitude of the observation errors (*Justification*). We provide simple formulae for deriving the posterior distribution and relevant summaries (*Simplicity*). We show how the addition of this information materially reduces uncertainty in the measurand (*Benefit*) and we also show how the validity of the prior information can be

verified in practice (*Verification*). Our simple priors are proposed purely for this specific Type A evaluation. Since this evaluation is by far the most widely used model in laboratories around the world, the use of these simple tools to improve measurement results for this model has the potential to have a profound influence on grassroots metrology.

Section 3 considers the case where a measurement model has multiple inputs, in some or all of which it is possible to apply the suggested informative priors. The approach is illustrated in a numerical example, using a model with six inputs. The reduction in the standard uncertainty of the measurand achieved through using the two simple informative prior distributions is examined, and the validity of the prior information is tested. A second numerical example of a measurement model is also considered in which there are two input quantities, with Type A evaluation used for one and Type B for the other. Although it is very simple, the model is indicative of models used by many laboratories.

Section 4 summarizes the findings and conclusions of this article.

2. Type A evaluation

Although there are many forms of Type A uncertainty evaluation such as involving several quantities measured simultaneously and cases involving complex variables, the canonical and commonest example of Type A uncertainty evaluation in metrology is as follows. We have a sample $x = [x_1, \ldots, x_n]$ of n (real) observations, assumed to be distributed independently as $\mathrm{N}(\mu, \sigma^2)$, where μ is the (unknown) population mean, while σ^2 is the (unknown) population variance. In section 3, we will consider measurement models in which a measurand is expressed in terms of two or more input quantities but here, in the context of its Type A evaluation or measurement, we refer to μ as the measurand.

We denote the sample mean by $\bar{x}$ and the sample variance by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

2.1. Informative and NIP distributions

The standard non-Bayesian Type A evaluation for this problem, given in the GUM, estimates μ by $\bar{x}$ and expresses uncertainty through a standard uncertainty $u(\bar{x})$ and expanded uncertainty $U(\bar{x})$ given by

$$u(\bar{x}) = \frac{s}{\sqrt{n}}, \qquad U(\bar{x}) = k(n-1)\frac{s}{\sqrt{n}},$$

where $k(d)$ is the 97.5 % quantile of the Student's t distribution with d degrees of freedom. (We use d to denote degrees of freedom rather than the more conventional ν because of the possible confusion with v used here for variance.)

JCGM 101 [3] gives a Bayesian Type A evaluation for this problem in which the posterior distribution of μ is a scaled and shifted t distribution with mean $\bar{x}$, standard deviation

$$u(\mu) = \sqrt{\frac{n-1}{n-3}}\,\frac{s}{\sqrt{n}}$$

and degrees of freedom $n-1$.

Part of the controversy over frequentist versus Bayesian inference in metrology concerns the difference between the two standard uncertainties. The Bayesian $u(\mu)$ is larger than the frequentist $u(\bar{x})$: the former is $\sqrt{(n-1)/(n-3)}$ times the latter. However, the interval $\bar{x} \pm k(n-1)s/\sqrt{n}$ is a 95 % coverage interval for μ in both analyses.

The posterior distribution of μ given therein derives from a NIP distribution and is commonly used to represent no prior knowledge about either μ or σ^2. See appendix B for details of this distribution.

The metrologist may of course have, on the basis of experience or judgment, subjective prior knowledge about the measurand, which can be formally incorporated into the analysis (as for instance in [15, 16]), but it will frequently be thought controversial, inappropriate or even inadmissible to influence the estimated value of the measurand in this way. However, relatively uncontroversial prior information about σ^2 will very often exist. The incorporation of such information in the Type A evaluation through the use of an informative prior distribution should be reflected in reduced posterior uncertainty. Specifically, a reduction in standard uncertainty, would allow the metrologist to report a stronger measurement result.

Prior information to the metrologist often relates to the observation error variance obtained with a well-characterized measuring system, that is, a system possessing stated measurement repeatability (measurement precision under a set of repeatable conditions of measurement [1, definition 2.21]). It is not so much a question of location of the mean of a set of observations, but of their standard deviation or some other measure of spread that is most relevant. A measuring system, such as a force-measuring machine, may have consistent repeatability over a certain force interval and be required to be capable of measuring any force within that interval. Presented with an unknown force, we may have meagre knowledge of the force itself but reasonable prior knowledge based on routine use of the machine of the standard deviation of the observations made of that force.

The GUM recognizes such a situation, stating in [2, clause 4.2.4]:

> 'For a well-characterized measurement under statistical control, a combined or pooled estimate of variance s_p^2 (or a pooled experimental standard deviation s_p) that characterizes the measurement may be available. In such cases, when the value of a measurand q is determined from n independent observations $[q_k, k = 1, \ldots, n]$, the experimental variance of the arithmetic mean $\bar{q}$ of the observations is estimated better by s_p^2/n than by $s^2(q_k)/n$ and the standard uncertainty is $u = s_\mathrm{p}/\sqrt{n}\ldots$'

However, in frequentist statistics, there are just two extremes regarding a parameter: either it is completely unknown or it is known. GUM clause 4.2.4 is not Bayesian,

just an example of the latter. It is also suspect because it treats the variance as being known exactly (with the added assumption that that value, supposedly from previous applications of the same measurement process, applies exactly to the current application). The frequentist must ignore either the evidence from the data or else all prior knowledge. Our approach is Bayesian because it uses both, recognizing that the data are relevant because there is prior uncertainty about the variance.

The family of scaled inverse-chi-squared (ICS) distributions we present in section 2.2 is a standard tool for expressing uncertainty about a variance in Bayesian statistics generally, and has been presented for use in the particular context of Type A evaluation in [17]. The mildly informative prior (MIP) and strongly informative prior (SIP) distributions we propose are specific instances of the general ICS family, chosen with our target audience in mind to make accessing the benefits of prior information as simple as possible. The noninformative NIP prior distribution is also identified in appendix B as a limiting form of ICS distribution. We present clear and simple criteria for selecting MIP or SIP that we believe will be more straightforward to employ in a busy laboratory than the formulation in [17]. Furthermore, the reality checks we propose in section 2.6 to meet our Verification desideratum are possible only because restricting the choice of prior in practice to either MIP or SIP allows the necessary accumulation of evidence.

2.2. Simple informative prior distributions

In accordance with section 2.1, we now suppose that the metrologist can specify prior information about the error variance σ^2, but that the prior distribution for μ is to be noninformative, reflecting no prior knowledge about the value of the measurand. We recognize that Bayesian statistical methods are unfamiliar to most metrologists and can be complex to apply. For any such method to be adopted in regular metrological practice, it must satisfy the desirable criteria listed in section 1.6.

In principle, to identify a prior distribution that accurately represents the metrologist's prior knowledge is a non-trivial task [18]. In practice, however, prior information regarding σ^2 can generally be represented adequately as in section 2.1 by a member of the ICS family of distributions. A considerable advantage of the choice of the scaled ICS distribution for σ^2 is that it retains the familiar Student's t distribution, as in the GUM [2], as the posterior for μ (see [16, 17], for instance). In the interests of providing simple, readily implemented procedures, we recommend the following two specific members of that family, representing different strengths of prior knowledge.

- The MIP distribution. The weaker of the two distributions is denoted by MIP(v) and is recommended when the metrologist can be confident that σ^2 will lie within a factor of 9 either side of an estimate v, that is, between $v/9$ and $9v$.
- The SIP distribution. The stronger of the two distributions is denoted by SIP(v) and is recommended when the metrologist can be confident that σ^2 will lie within a factor of 3 either side of an estimate v, that is, between $v/3$ and $3v$.

Details of the ICS distributions, and of the MIP and SIP distributions in particular, are given in appendix B. In each case, the use of the distribution requires only a prior estimate v of σ^2. The choice of distribution, MIP or SIP, expresses the strength of prior knowledge, through a judgment of confidence that σ^2 will lie within a factor 9 or 3, respectively, either side of v. Specifically, the metrologist should feel at least 95 % certain that σ^2 will be in that interval.

Appendix B shows that these prior distributions are easy to specify and to justify in practice, meeting the *Justification* criterion. Furthermore, they meet the *Simplicity* criterion because they have the property that estimates and uncertainty measures can be derived in closed form as simple formulae.

The choice of prior distributions for use in metrology has been considered by several authors; see, for instance [16, 19, 20]. In particular, in [20] various applications are presented in which prior distributions are derived from historical information such as previous proficiency tests or interlaboratory comparisons. It is shown how these applications profit from the use of such knowledge. In contrast, we assume here that such knowledge is unavailable or that to use it would likely be beyond the capabilities of many metrologists in calibration or testing laboratories. All that is asked is that the metrologist makes a judgment of the factor within which the variance of the measurand is expected to lie. That judgment calls on some knowledge of measurement precision [21, definition 3.3.4] for the task in hand but does not entail anything more sophisticated than that.

The MIP distribution represents weak prior knowledge about a variance, and we suggest that at least this degree of prior knowledge will be justifiable in many problems in metrology. In a good proportion of these problems, the more informative SIP distribution should also be justifiable. Instances of the availability of prior information where these distributions could be applicable are in dimensional metrology [22], sludge, biowaste and soil sampling [23], and manufacturing metrology [24].

Other informative priors have been used for Type A uncertainty evaluation, but fail to satisfy our criteria. Cox and Shirono [25] use a Jeffreys' prior truncated above and below. Truncating the Jeffreys' prior either involves an arbitrary choice of truncation points or else explicit judgments about where to place them, knowing that if the resulting prior is to have any informative value the answer will depend on those points. Such choices will be challenging to justify in practice, and the truncated distribution does not yield simple formulae for the estimate or uncertainty measures. Van der Veen [26] considers some weakly informative priors for various forms of Type A uncertainty evaluation, but there are again no simple formulae for the estimate or measures of uncertainty. Instead,

calculations need to be carried out using Markov chain Monte Carlo methods.

In a documentary standard concerned with sampling plans [27], the level of prior information is specified on an ordinal scale 'Trust', by choosing between low, mid and high. Trust level low is used if no prior experience with populations submitted for inspection exists, high if there is strong evidence of good performance, and mid if there is weak evidence of good performance. Beta distributions (without motivation) are used as priors with the two parameters a and b of those distributions depending on the Trust level. A rectangular distribution (with $a = b = 1$) is used for Trust level low. Advice is given on the choice of parameter values for Trust levels mid and high (for which choosing $a \leqslant 1$ and $b \geqslant 1$ but not $a = b = 1$) results in a strictly decreasing beta curve. Although used in a different context, the Trust levels low, medium and high, respectively, can be contrasted with the NIP, MIP and SIP priors used here.

2.3. The effect of additional information

A proposed new method should of course also offer some tangible advantage over an existing method (*Benefit*), which in this case should be an expected reduction in uncertainty measures. The existing method [3] is Bayesian analysis with NIP distribution in section 2.1. We will compare uncertainty measures obtained using the NIP distribution with those obtained with a MIP or SIP distribution.

The most widely used measure of uncertainty in metrology is the standard uncertainty, defined in the GUM [2] as a standard deviation. However, comparison on the basis of a standard deviation is problematic because (a) it has conceptually different interpretations in the frequentist and Bayesian paradigms, and (b) it can be infinite, being inapplicable for samples of size less than four.

To compare uncertainties and coverage intervals for the priors considered, we use *characteristic uncertainty* $c(\mu)$ [28], defined as one quarter of the length of a 95 % coverage interval. Thus, $c(\mu)$ expresses uncertainty in a more direct and meaningful way than a standard deviation. Furthermore, unlike standard deviation, the characteristic uncertainty exists for any probability distribution. Moreover, it is appropriate to report characteristic uncertainty in our numerical examples below since considerations of interpretation of coverage intervals according to the frequentist and Bayesian paradigms are totally avoided. These and other arguments in favour of characteristic uncertainty as a preferred uncertainty measure are presented in [28]. Metrologists who nevertheless prefer to work with and report standard uncertainties will find appropriate details in appendix B.4. Conclusions presented here in terms of characteristic uncertainty will hold equally strongly for standard uncertainties.

Formally, the characteristic uncertainty is defined by reference to the median estimate, which is advocated by [28] as a more meaningful estimate of a measurand than the mean. However, in this work we will mostly be concerned with Student's t distributions, for which the mean and median are identical.

Table 1. Posterior degrees of freedom d^* and variance estimates v^* for three prior distributions.

Prior	d^*	v^*
NIP	$n-1$	s^2
MIP(v)	$n+2$	$\dfrac{3v+(n-1)s^2}{n+2}$
SIP(v)	$n+7$	$\dfrac{8v+(n-1)s^2}{n+7}$

Appendix B.3 shows that for all three prior distributions the posterior distribution of μ is a scaled and shifted Student's t distribution with mean $\bar{x}$. Hence all three distributions yield the same estimate of μ,

$$m(\mu) = \bar{x}.$$

However, they yield different characteristic uncertainties

$$c(\mu) = \frac{k(d^*)}{2}\sqrt{\frac{v^*}{n}}. \tag{1}$$

with the values of d^* and v^* given in table 1.

In effect, the prior information in the MIP and SIP distributions is equivalent to a pseudo-sample of 4 or 9 additional observations, respectively, in each case with a pseudo-sample variance of v. In the d^* column of table 1, we see that the additional pseudo-sample size increases the sample degrees of freedom $n-1$ by 3 or 8, respectively. In the v^* column, the pseudo-sample variance v is combined with the sample variance s^2 in the natural way to produce the revised estimate v^*.

The effect of these informative prior distributions on the characteristic uncertainty is seen primarily in the posterior degrees of freedom d^*. Increasing d^* reduces the factor $k(d^*)$. Moving from NIP to MIP to SIP produces systematic reductions in these uncertainty factors.

The effect of the prior information on the v^* term is to pull the sample estimate s^2 of σ^2 towards the prior estimate v. The pull is stronger with the more informative SIP than with MIP. If s^2 is larger than v, then v^* will be smaller than s^2, again reducing the characteristic uncertainty. Conversely, if s^2 is smaller than v, the uncertainty will be increased. However, v^* is expected to be neither larger nor smaller than s^2; they are both estimates of σ^2. Overall, therefore, the informative prior distributions will reduce uncertainty, primarily by increasing the posterior degrees of freedom d^*.

2.4. A basic algorithm

The procedure for conducting a Type A evaluation with a MIP or SIP prior distribution is simply set out in the following algorithm. In this algorithm, the reporting stage follows [28] with (a) a summary report comprising the median estimate and characteristic uncertainty, and (b) a full report giving the distribution. Algorithms for reporting based on standard uncertainties are given in appendix B.4.

Algorithm 1. Type A evaluation from a normal sample with MIP or SIP prior information.

Begin:
Given:
- A sample $x_1, \ldots, x_n$ of n indications, normally distributed with mean μ the quantity of interest and unknown variance σ^2;
- A prior estimate v of σ^2 such that the metrologist can be confident that
  (MIP) σ^2 will lie between $v/9$ and $9v$, or
  (SIP) σ^2 will lie between $v/3$ and $3v$

Compute:
1. Form sample mean $\bar{x} = \sum_1^n x_1/n$
2. Form sample sum of squares $S = \sum_i^n (x_i - \bar{x})^2$
3. Set prior degrees of freedom
   (MIP) $d = 3$
   (SIP) $d = 8$
4. Form posterior degrees of freedom $d^* = n + d - 1$
5. Form posterior variance estimate $v^* = (dv + S)/d^*$
6. Form characteristic uncertainty $c(\mu) = k(d^*)\sqrt{v^*}/(2\sqrt{n})$

Report:
(a) The measured value (median estimate) of μ is $\bar{x}$ with characteristic uncertainty $c(\mu)$
(b) The quantity μ has a t distribution with mean $\bar{x}$, scale parameter v^*/n and degrees of freedom d^*

End

2.5. Advantages of the proposed prior distributions

To illustrate the advantages of the informative MIP and SIP prior distributions over the noninformative NIP distribution, as well as to identify the price to be paid for those advantages, we present a simple example in the case of one normal sample. We contrast the characteristic uncertainties and coverage probabilities from the noninformative NIP prior distribution, with those from the informative MIP and SIP distributions. We set the sample size to $n = 5$ and for MIP and SIP the prior estimate of the variance to $v = 1$.

Figure 1 (left) shows the expected characteristic uncertainty as a function of σ using the NIP, MIP and SIP prior distributions. The expected characteristic uncertainty using the NIP distribution, which is equivalent to the standard frequentist analysis of the GUM, is linear in σ, while the values for MIP and SIP, computed by a Monte Carlo method with 10^6 samples as set out in appendix C.1, show the influence of the prior information.

For the SIP prior distribution, the metrologist's judgment is that σ^2 is highly likely to lie within a factor 3 either side of the prior estimate $v = 1$, and hence that σ is highly likely to be in the interval $3^{-1/2} = 0.577$ to $3^{1/2} = 1.732$. This interval is shown as the green horizontal bar in figure 1. The prior distribution will yield a characteristic uncertainty that is larger on average than the standard GUM proposal (the NIP line in figure 1) when σ is smaller than the metrologist expected. Conversely it will give a lower characteristic uncertainty when σ is as the metrologist expected or larger. Overall, if σ^2 were indeed drawn randomly according to the metrologist's SIP prior distribution then with probability 0.9 this prior distribution would give a lower characteristic uncertainty than the NIP distribution; the median percentage reduction (the reduction achieved or exceeded with probability 0.5) would be 19.1 %.

Turning to the weaker MIP prior distribution, the metrologist's judgment in this case is that σ will lie in the interval 1/3 to 3 with high probability. This interval is shown as the red horizontal bar in figure 1. We see a similar pattern to that observed with the SIP distribution. Overall, if σ^2 were indeed drawn randomly according to the metrologist's MIP prior distribution, this prior distribution would give a lower characteristic uncertainty than the NIP distribution with probability 0.8, and the median percentage reduction would be 15.9 %.

These gains relative to the standard GUM formulation are achieved conditional on the metrologist's prior information being valid, in the sense that over many applications the true values of the underlying error variance σ^2 behave as if drawn randomly from the stated prior distribution. To assess the consequences of the prior information not being valid in this sense, we consider the coverage of the implied 95 % coverage interval $m(\mu) \pm 2c(\mu)$. Figure 1 (right) shows the coverage probability as a function of σ for the three priors, and again the most likely ranges for σ according to the MIP and SIP prior distributions are shown.

The coverage probability is identically 0.95 (95 %) by construction for the NIP distribution, for all σ^2. For both informative priors, the expected coverage is also 95 % if the prior distribution is valid. However, figure 1 (right) shows the consequence if the metrologist misjudges the likely values of σ^2. The coverage in both cases is a decreasing function of σ and for small values of σ approaches 100 %. Therefore, if the true error variance is appreciably smaller than the metrologist expects, then the informative prior will result in a conservative evaluation of uncertainty. That is, the metrologist will report a characteristic uncertainty implying a 95 % coverage interval that in fact is almost certain to contain the true value of μ.

Conversely, though, if the metrologist underestimates the likely value of σ^2, the resulting characteristic uncertainty will also be underestimated and the implied 95 % coverage interval will have an actual coverage probability appreciably below 95 %. If, for instance, despite specifying the SIP(1) prior distribution, and hence that the metrologist is confident that σ will not exceed 1.732, the true value of σ is 3 or more, then the coverage probability will be less than 80 %. According to the metrologist's specified SIP(1) distribution, $\sigma \geqslant 3$ has probability just 0.0012, so the risk of such poor coverage is extremely low and would be of no concern. It becomes a concern only if the metrologist has mis-specified the prior distribution by seriously underestimating σ^2, or by over-stating their confidence by using the SIP(1) prior when only MIP(1) is justified. Just as the gains in terms of reduced characteristic uncertainty are greater with the stronger SIP prior distribution, the consequences of mis-specifying the prior information, and in particular of under-estimating σ^2, are greater. It is noted that in many papers the consequences of mis-specifying the prior information are overlooked.

This is why we have emphasized that prior distributions must represent honest and justifiable prior knowledge.

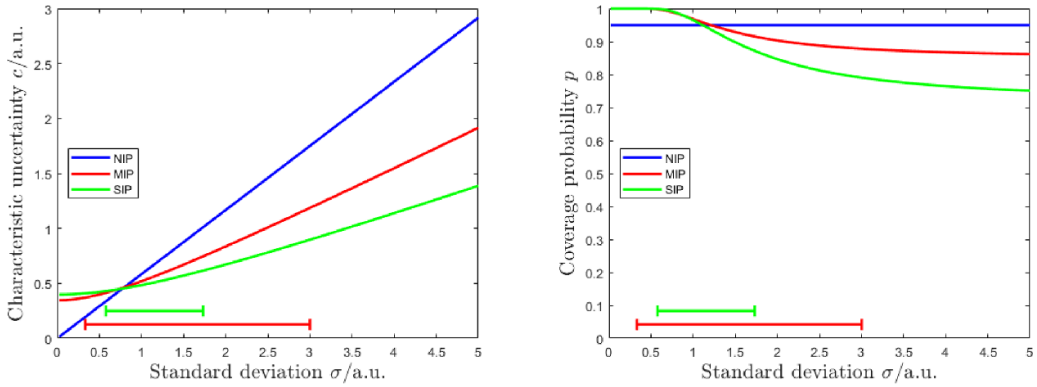

Figure 1. Expected characteristic uncertainty and (right) coverage probability as a function of σ using the NIP (noninformative), MIP (mildly informative) and SIP (strongly informative) prior distributions for a sample of size 5.

Table 2. Median percentage reduction in characteristic uncertainty using MIP and SIP prior distributions, relative to the noninformative NIP distribution, for various sample sizes n.

n	2	3	4	5	6	7	8	9	10
MIP/%	72.9	37.9	23.1	15.9	11.9	9.3	7.6	6.4	5.5
SIP/%	75.5	42.0	26.9	19.1	14.6	11.6	9.5	8.1	7.0

The above illustrations are for $v = 1$, but the advantages of utilizing MIP and SIP prior distributions would be the same for any value of v if they were valid judgments, and the consequences of under-estimating σ^2 would be the same for any value of v. In particular, graphs of expected characteristic uncertainty and coverage probability for different values of v would be identical to figure 1 (left and right, respectively) except that the x-axis values would be multiplied by $\sqrt{v}$.

The illustrations are also for $n = 5$. The reduction in characteristic uncertainty will be smaller for larger values of n, but will also be even larger for $n < 5$, as shown in table 2.

We suggest that for sample sizes of 8 or less, the informative prior distributions satisfy our criterion of *Benefit* by offering substantial reductions in reported uncertainty.

2.6. *Reality checking*

As we have stressed, there is an onus on the metrologist to justify the choice of an informative prior distribution. The primary requirement is to be able to document the evidence and experience in support of a prior estimate v for σ^2 such that the metrologist can be confident that σ^2 lies within a factor 3 (for SIP) or 9 (for MIP) either side of v. The sample data cannot be used to estimate or suggest a value for v, since to do so (such as by setting $v = s^2$) would entail counting the data twice; the evidence in support of the prior distribution must be *prior* knowledge. Two suggestions can be made to assist with the choice.

First, it is safe to err on the side of expressing weaker prior information. If, for instance, the metrologist can justify confidence that σ^2 lies within a factor 5 of v, then the MIP prior is acceptable. The expected coverage of the resulting 95 % intervals with the weaker prior will actually be greater than 95 %,

Table 3. Percentiles of F distributions.

	MIP				SIP			
n	25th	50th	75th	95th	25th	50th	75th	95th
2	0.12	0.59	2.02	10.13	0.11	0.50	1.54	5.32
3	0.32	0.88	2.28	9.55	0.30	0.76	1.66	4.46
4	0.42	1.00	2.36	9.28	0.41	0.86	1.67	4.07
5	0.49	1.06	2.39	9.12	0.49	0.91	1.66	3.84
6	0.53	1.10	2.41	9.01	0.53	0.95	1.66	3.69
7	0.56	1.13	2.42	8.94	0.56	0.97	1.65	3.58
8	0.58	1.15	2.43	8.89	0.59	0.99	1.64	3.50
9	0.60	1.16	2.44	8.85	0.61	1.00	1.64	3.44
10	0.61	1.17	2.44	8.81	0.63	1.01	1.63	3.39

while there will still be some gain in terms of shorter intervals and reduced characteristic uncertainty compared with the standard frequentist method of the GUM.

Second, it is possible to test whether the observed sample variance s^2 is consistent with the stated prior distribution for σ^2. Formally, if the prior distribution is valid, the predictive distribution of the ratio s^2/v is shown in appendix B.5 to be $\mathrm{F}_{n-1,8}$ for the SIP prior, and $\mathrm{F}_{n-1,3}$ for the MIP prior, where F_{ν_1,ν_2} denotes the Snedecor F distribution with degrees of freedom ν_1 and ν_2. There would be concern about the validity of the prior distribution if the ratio is very large, since this would suggest that σ^2 is larger than expected by the prior distribution, and this is when the coverage decreases significantly. It is therefore suggested that a laboratory using these simple informative prior distributions should routinely compute s^2/v. Over time, these values should resemble random draws from the corresponding F distribution. A single particularly large value should be cause for investigation and possible choice of an alternative or weaker prior distribution.

The metrologist does not, however, need to be familiar with F distributions because table 3 suffices to facilitate these checks. For each value of n from 2 to 10, and for each informative prior distribution, four percentiles are given, the 25th, 50th, 75th and 95th. Over time, when using these distributions for many measurements, the metrologist should find that approximately equal numbers of the computed s^2/v values lie below the corresponding 25th percentile, between the 25th and 50th,

between the 50th and 75th, and above the 75th. Values above the 95th percentile should be found only occasionally (approximately once in 20 measurements).

If the four proportions are far from equal, or if values of s^2/v exceed the 95th percentile, the following actions are suggested.

- If many more s^2/v values fall below the corresponding 50th percentiles than above, the metrologist may be tending to give values of v that are too large, thereby overestimating the values of σ^2. Smaller characteristic uncertainty values could overall have been reported by better prior estimation of σ^2.
- Conversely, if many more s^2/v values fall above the corresponding 50th percentiles than below, the metrologist may be tending to give values of v that are too small, thereby underestimating the values of σ^2. This is a more serious departure from the norm of equal proportions, since it will mean that the metrologist's reported characteristic uncertainty values may have been overall too small.
- If, when using the MIP prior distribution, many more s^2/v values fall between the corresponding 25th and 75th percentiles than outside this interval, the metrologist could more often justify using the stronger SIP distribution.
- Conversely, if when using the SIP prior distribution many fewer s^2/v values fall between the corresponding 25th and 75th percentiles than outside that interval, the metrologist is often using the stronger prior distribution when only the weaker MIP distribution would be justified.
- A single value of s^2/v exceeding the 95th percentile is a cause for concern because it suggests that the informative prior distribution may not be valid, and that the characteristic uncertainty is likely to be underestimated. Such values can be expected to occur by chance, about once in every twenty measurements, even if the prior distribution is valid but should always cause metrologists to check their justification.

We suggest that table 3 and the above check actions could be provided as a standard laboratory reference document. Quality assurance procedures are familiar requirements in any laboratory, and what we are proposing here is in that sense nothing new, and indeed to be expected.

The ability to validate the prior distribution by checking its consistency with the data is an important practical feature of our proposed simple informative Bayesian methods, and satisfies our *Verification* criterion.

3. Informative prior distributions for model inputs

In section 2 we considered a single Type A evaluation of a measurand X, but measurement often involves a measurement model to relate the measurand Y to a number of inputs $X_1, X_2, \ldots$, each of which may have a Type A or Type B evaluation. The simple informative prior distributions proposed in section 2.2 have a role to play here, too, since reduced uncertainty about any of the model inputs should lead to reduced uncertainty about the measurand. We will illustrate the extent of this reduction using an example in which there are six Type A evaluations.

We follow that example by an example involving a model with a single Type A and a single Type B evaluation. Such a model is commonly employed in testing laboratories where n may be as small as 3.

Additional digits in the results are reported for comparison purposes.

3.1. A Monte Carlo algorithm

Our examples will employ the Monte Carlo method of [3] to compute the characteristic uncertainty of a measurand defined by a measurement model. The Monte Carlo method allows arbitrarily accurate propagation of uncertainty in measurement models. Approximate propagation methods, referred to as the GUM uncertainty framework and the characteristic uncertainty framework, are contrasted in [28].

Algorithm 2. Propagating uncertainty through a measurement model by the Monte Carlo method.

Begin:
Given:
- A measurement model $Y = f(\boldsymbol{X})$ expressing a measurand Y in terms of N independent inputs $\boldsymbol{X} = (X_1, \ldots, X_N)$;
- A probability distribution for each of the inputs;
- A size M for the Monte Carlo sample.

Compute:
1. For each input X_i, draw M random values $X_{i1}, \ldots, X_{iM}$ from its probability distribution
2. Form M input samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$, where $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{Nj})$
3. Form the measurand sample $Y_1 = f(\boldsymbol{X}_1), \ldots, Y_N = f(\boldsymbol{X}_M)$
4. Arrange the measurand sample in increasing order $Y_{[1]} \leqslant \ldots \leqslant Y_{[M]}$ and form the median:
   (M odd) $m(Y) = Y_{[(M+1)/2]}$
   (M even) $m(Y) = (Y_{[M/2]} + Y_{[1+M/2]})/2$
5. Form the characteristic uncertainty $c(Y)$ as the smallest positive value such that the interval $[m(Y) - 2c(Y),\ m(Y) + 2c(Y)]$ contains 95 % of the measurand sample

Report:
(a) The measured value (median) of Y is $m(Y)$ with characteristic uncertainty $c(Y)$.
(b) The probability distribution of Y is represented by the discrete distribution with M values $Y_1, \ldots, Y_N$, each having probability $1/M$.

End

Step 1 of the *Compute* stage in algorithm 2 requires random samples to be drawn from the probability distributions of the inputs. Procedures for sampling directly from standard probability distributions are available in many computing packages. In particular, scaled and shifted t distributions may be available to sample directly. If not, they can be readily sampled indirectly if only standard Student's t distributions are available. For example, if t is a randomly sampled value from

a Student's t distribution with d degrees of freedom, then

$$x = m + \sqrt{w}\,t$$

will be a randomly sampled value from the scaled and shifted t distribution with mean m, scale parameter w and degrees of freedom d.

The *Report* stage of algorithm 2 reports the median and the characteristic uncertainty because, as we have argued in [28], we believe these are more meaningful as an estimate and an expression of uncertainty for recipients of a measurement report. However it is simple to modify steps 4 and 5 of the *Compute* stage to compute the more traditional mean and standard deviation.

3.2. Numerical example 1: single burning item test

Measurement of the rate of oxygen consumption constitutes a versatile and powerful tool for estimating the rate of heat release in fire experiments and fire tests [29, 30]. Since the heats of combustion per unit of oxygen consumed are approximately the same for most fuels commonly encountered in fires, a measured rate of oxygen consumption can be converted to a reliable measure of heat release rate. Standards Publication CEN/TR 16988:2016 [31] provides a method for estimating the velocity and associated uncertainty of flows associated with small to medium size fires. A velocity-pressure probe, a device relating the velocity of the exhaust gases to differential pressure, measures the volume flow through an exhaust duct. The so-called flow profile correction factor, denoted here by κ, converts the velocity at the axis of the probe to the mean velocity over the cross-section of the duct. It is directly proportional to the volume flow and therefore to the heat release rate. In [31, clause 2.5.13.2], κ is expressed by the model

$$\kappa = \frac{1}{5}\sum_{i=1}^{5}\frac{w_i}{w_c} \qquad (2)$$

with six input quantities. In expression (2), w_i, $i = 1,\ldots,5$, are measurements taken on five different radii and w_c is a central measurement. Each measurement is actually the average of four independent indications taken at 90° intervals. The six GUM Type A evaluations are reported in table 4. The characteristic uncertainty of each input is the standard uncertainty multiplied by $k(3)/2 = 1.591$. The same results would be obtained from Bayesian Type A evaluations using the non-informative NIP prior distribution in each case.

Cox and O'Hagan [28] propagate these uncertainties through the model (2) using the Monte Carlo method, showing that the median estimate of κ is 0.817 and its characteristic uncertainty is 0.076. They also employ an approximate computation in which the law of propagation of uncertainty is used to propagate characteristic uncertainties through a linearized version of the model, obtaining the same median estimate and an approximate characteristic uncertainty of 0.075. We now consider the effect of introducing informative prior distributions.

We suppose that the metrologist provides the same prior distribution for each input's σ^2 parameter, with an estimate of $v = 1\,\mathrm{m^2\,s^{-2}}$. That is, in each case the metrologist estimates the standard deviation of the Gaussian sampling error to be $1\,\mathrm{m\,s^{-1}}$. (Note that we are not assuming that the inputs have the same σ^2 parameter. The parameters only have the same, independent, prior distributions, thereby allowing them to differ within the range of that common prior distribution.) We consider the effect of using independent MIP(v) prior distributions for all six inputs, and of using independent SIP(v) prior distributions instead.

The evaluation of each input can now be made by following algorithm 1 in section 2.4. The sum of squares S required in the algorithm can be inferred by noting that $S = (n-1)s^2 = n(n-1)u^2(w_i)$. The $u(w_i)$ values are given in the third column of table 4 and in each case $n = 4$. The evaluations yield scaled and shifted t distributions with degrees of freedom $d^* = 6$ and $d^* = 11$ with MIP and SIP prior distributions, respectively. In table 5, the second and third columns give the corresponding values of v^*. Using the Monte Carlo algorithm 2, we sample from these distributions and apply expression (2) to obtain a sample of κ, from which the characteristic uncertainties are found to be 0.052 with MIP prior distributions and 0.045 with SIP distributions. More details of these computations can be found in appendix C.2.

As expected, the additional prior information has reduced the measurement uncertainty regarding κ compared with the characteristic uncertainty of 0.076 obtained with NIP distributions. The reductions achieved by the MIP and SIP distributions are substantial in this example, more than 30 % and 40 % respectively.

We now apply the reality check suggested in section 2.6. The relevant row of table 3 is for $n = 4$. We have $v = 1$ and for each quantity the relevant value of s^2 is 4 times the square of the standard uncertainty in the third column of table 4. We only have six values for s^2/v, but none of the checks suggested there indicates a problem with the prior distribution except the last one. If we use SIP prior distributions then one of the values (5.13) exceeds the 95th percentile (4.07). Although one such instance in six measurements is not particularly unexpected, the reality checks suggest that in this case the metrologist should use the MIP distribution.

The prior information in this example is of course not a genuine metrologist's opinion but arbitrarily chosen for the purpose of illustration. In practice, metrologists having such a sample of just six s^2/v values might still use the SIP prior distribution if they felt it could be justified.

3.3. Numerical example 2: model involving a Type A and a Type B evaluation

In [28], consideration was given to the measurement model

$$Y = X_1 + X_2,$$

where the measurand Y is modelled as a quantity X_1, evaluated as the sample mean of n independent normally distributed observations, plus an independent correction term X_2.

Table 4. GUM/NIP evaluations for the single burning item test.

Quantity	Estimate/ms^{-1}	Standard uncertainty/ms^{-1}	Degrees of freedom	Characteristic uncertainty/ms^{-1}
w_1	7.00	1.132	3	1.801
w_2	9.39	0.412	3	0.656
w_3	10.62	0.531	3	0.845
w_4	11.25	0.180	3	0.286
w_5	12.37	0.233	3	0.371
w_c	12.39	0.636	3	1.012

Table 5. Computations with MIP and SIP prior distributions, Example 2.

Input quantity	MIP v^*	SIP v^*	s^2/v
w_1	3.0628	2.1252	5.13
w_2	0.8395	0.9124	0.68
w_3	1.0639	1.0349	1.13
w_4	0.5648	0.7626	0.13
w_5	0.6086	0.7865	0.22
w_c	1.3090	1.1685	1.62

Suppose the measurand is the primary length of some artefact of nominal length 100 m and $n = 5$ independent normally distributed observations

$$(99.87,\ 99.58,\ 99.52,\ 99.93,\ 99.60)\ \text{m}$$

of X_1 are made, with a sample mean of 99.700 m. Reporting as in [28], X_1 has median 99.700 m, standard uncertainty 0.083 m and characteristic uncertainty

$$c(X_1) = u(x_1)k_{n-1}/2 = 0.116\,\text{m}.$$

Various sources of information regarding X_2 were considered in [28]. Here, we select one of these sources: a Type B evaluation is carried out for X_2 assuming that the correction lies between -0.10 m and 0.10 m. A uniform (rectangular) distribution is assigned between these bounds, which therefore has mean and median 0.000 m, and characteristic uncertainty

$$c(X_2) = 0.95/2 \times 0.10\,\text{m} = 0.048\,\text{m}.$$

It was judged that for X_1 a prior estimate v of the variance σ^2 of the MIP and SIP was 0.02 m^2. Applying algorithm 1 in section 2.4, gave posterior degrees of freedom $d^* = 7$ and posterior variance estimate $v^* = 0.028\,\text{m}^2$ for the MIP. The corresponding values for the SIP were $d^* = 12$ and $v^* = 0.025\,\text{m}^2$.

The Monte Carlo algorithm 2 was applied with $M = 10^7$ trials, with the uniform distribution for X_2 and the t distribution for X_1 from algorithm 1. The resulting characteristic uncertainty of Y was 0.112 m for the MIP and 0.106 m for the SIP. These compare with 0.127 m for the NIP (essentially the method in [3]). The MIP gave a 12 % reduction over NIP and the SIP a 17 % reduction.

The reality check of section 2.6 gave $s^2/v = 1.73$, considerably smaller than the 95th percentiles of the F distribution for $n = 5$.

4. Conclusions

The metrologist frequently has prior knowledge concerning the likely magnitude of errors in the sample indications for a quantity subject to Type A evaluation. We have presented two simple prior distributions, the MIP and SIP distributions, to encode such prior information about the observation error variance. We have shown how they can be rigorously justified in practice through specific prior judgments, presented simple formulae to incorporate them in a Bayesian Type A evaluation and given equally simple procedures to verify their validity over a series of measurements. Explicit algorithms have been provided for the Type A evaluation of a single measurand, and for the case when it is an input to a measurement model. We have presented examples illustrating the advantages of these prior distributions in terms of reduced measurement uncertainty in Type A evaluation, showing how even larger reductions in uncertainty for a measurand can be achieved when several Type A evaluations contribute to a measurement model. Although those reductions have been presented in terms of characteristic uncertainty, for the reason stated in section 2.3, almost identical reductions are achieved in standard uncertainty.

The mildly informative MIP and the more strongly informative SIP prior distributions have thereby been shown to satisfy the desiderata of *Justification, Simplicity, Benefit and Verification* for informative Bayesian methods to be acceptable for use in metrology.

Type A evaluation from a sample of indications assumed to be normally distributed is employed daily by metrologists in laboratories worldwide. In all these applications, the MIP and SIP prior distributions would offer substantial reductions in measurement uncertainty over the existing GUM procedure, without requiring any more sophisticated computations.

It is our hope that the manifest benefits to their work, and to their clients, from the adoption of these simple informative priors will stimulate our target audience to seek similarly simple and effective Bayesian methods in other common Type A evaluations. Further, we hope that researchers will meet that demand by adapting more complex Bayesian methods to the practicalities of everyday metrology.

Acknowledgments

This work was supported by an ISCF (Industrial Strategy Challenge Fund) Metrology Fellowship grant provided by the UK

Government's Department for Business, Energy and Industrial Strategy (BEIS). It was also supported by BEIS as part of NPL's Data Science programme. The authors benefited greatly from discussions with Alistair Forbes (NPL) and members of the Joint Committee for Guides in Metrology (JCGM) although the views expressed are those of the authors. Moreover, the referees made many helpful comments.

Appendix A. Expectations of Bayesian coverage intervals

Here we prove the result stated in section 1.4 concerning the expected coverage of Bayesian intervals.

In this appendix we denote the measurand by θ and the data to be used in a Type A evaluation of θ by t. Bayesian methods derive the posterior distribution of θ, denoted by $p(\theta \mid t)$ by applying Bayes' theorem to combine the prior distribution $p(\theta)$ with the information in the data, represented by the *likelihood function* $p(t \mid \theta)$. Various forms of Bayesian inference may be obtained from the posterior distribution. We focus on a 95 % coverage interval, usually referred to in Bayesian analysis as a *posterior credible interval* $\Theta(t)$ defined such that $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$, a conditional probability that applies for given data t. The claim in section 1.4 is that the unconditional probability $\mathrm{P}(\theta \in \Theta(t))$ is also 0.95.

There are two random variables here, θ and t, and we consider an event, $\theta \in \Theta(t)$, that depends on both. In general, consider an event F depending on two random variables Y and Z. A standard result in probability theory (the law of total probability, a special case of the law of iterated expectation) states

$$\mathrm{P}(F) = \mathrm{E}(\mathrm{P}(F \mid Z)). \tag{3}$$

The interpretation here is that on the left-hand side the probability of F is unconditional, and therefore averaged over the joint distribution of both Y and Z. On the right-hand side, the term $\mathrm{P}(F \mid Z)$ is the conditional probability of F, averaged over the conditional distribution of Y given Z. This conditional probability is in general a function of Z, and we then take the expectation of this function, averaging with respect to the marginal distribution of Z.

We will apply the general result in two ways. In both cases, we take F to be the event $\theta \in \Theta(t)$. First, let Y be the measurand θ and Z the data t; then the theorem says

$$\mathrm{P}(\theta \in \Theta(t)) = \mathrm{E}(\mathrm{P}(\theta \in \Theta(t) \mid t)),$$

but the Bayesian interval has the property that $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$, a constant, for all t, and the expectation of a constant is a constant. Therefore, the unconditional probability $\mathrm{P}(\theta \in \Theta(t))$ is also 0.95.

However, it should be recognized that the Bayesian posterior distribution is only a valid opinion of the metrologist regarding the measurand after seeing the data t if the prior distribution is a valid opinion before the data. Hence the statement $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$ and the above proof depends on the validity of the prior distribution.

The role of the prior distribution becomes clearer if we reverse the roles of Y and Z in expression (3), so that now Y is t and Z is θ. The theorem now says that

$$\mathrm{P}(\theta \in \Theta(t)) = \mathrm{E}(\mathrm{P}(\theta \in \Theta(t) \mid \theta)).$$

The probability $\mathrm{P}(\theta \in \Theta(t) \mid \theta)$ is the frequentist coverage probability, in which we consider the measurand θ to be fixed and compute the frequency with which $\theta \in \Theta(t)$ over an infinite sequence of random draws of the data t. For a frequentist 95 % interval, this coverage is 0.95 for all θ, which, being constant, the unconditional probability is also 0.95. For a Bayesian interval, however, $\mathrm{P}(\theta \in \Theta(t) \mid \theta)$ depends on θ. We have proved that the unconditional probability is 0.95, and hence its frequentist coverage will be 0.95 when averaged with respect to the prior distribution. The practical interpretation is that over a long sequence of measurements the Bayesian intervals will contain the true measurand values 95 % of the time only if the corresponding θ values behave as if sampled from the metrologist's prior distribution.

Appendix B. Simple informative prior distributions

The inverse-chi-squared (ICS) family of distributions is widely used in Bayesian statistics to represent prior information about a variance for normally distributed data. Applications of ICS distributions in metrology include stiffness of anisotropic solid materials using ultrasound spectroscopy [32], lithographic control using critical dimension scanning electron microscope [33] and construction of a curve fitted to points in a two-dimensional coordinate system [34]. In order to address the needs and concerns of our target audience, and to meet our four desiderata of section 1.6, we have identified two specific members of that family that can readily be used and justified in routine laboratory applications. In this appendix, we present some standard theory of ICS distributions and develop the MIP and SIP cases that are presented in section 2.2 as simple 'default' choices for variances of measurement devices in metrology.

If a parameter z has the ICS distribution with degrees of freedom d and scale v, which we write as $\sigma^2 \sim v d \chi_d^{-2}$, then the density function of z has the form

$$f(z) = \frac{(vd/2)^{d/2}}{\Gamma(d/2)} z^{-1-d/2} \exp\left(-\frac{vd}{2z}\right).$$

The expectation $\mathrm{E}(z)$ and variance $\mathrm{Var}(z)$ of z are

$$\mathrm{E}(z) = \frac{d}{d-2}v, \qquad \mathrm{Var}(z) = \frac{2d^2}{(d-2)^2(d-4)}v^2 = \frac{2}{d-4}\mathrm{E}^2(z),$$

provided $d > 2$ and $d > 4$, respectively [35, section 11.5].

ICS distributions provide a flexible family to represent prior information about a variance parameter σ^2. They are defined for positive quantities, and through the choice of degrees of freedom and scale they can represent a wide range of prior

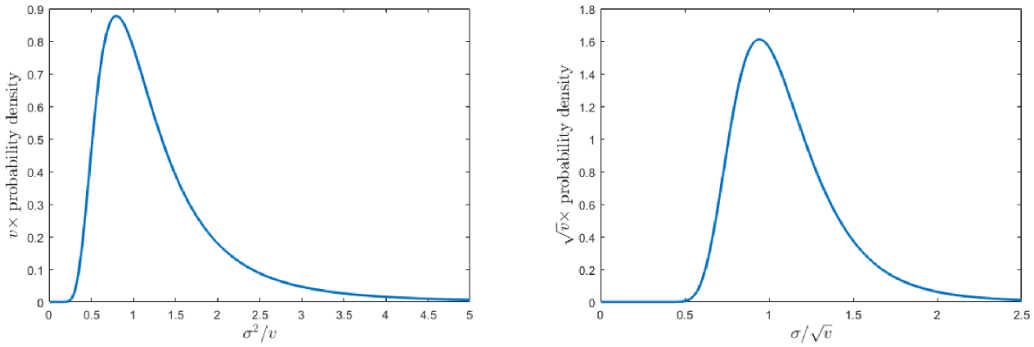

Figure 2. Density function of SIP(v) and (right) of the square root of SIP(v).

knowledge about σ^2. For instance, if the analyst has a prior mean t for σ^2, with variance w, then this information can be represented by an ICS prior distribution with degrees of freedom $d = 4 + 2t^2/w$ and scale $v = t(w + t^2)/(2w + t^2)$. Alternatively, [17] interprets $d + 1$ as a number of observations corresponding to the strength of prior information, and also suggests a more complex way of assigning d using a quantile judgment.

There are better ways to specify a prior distribution. Formal protocols for eliciting expert judgments, such as the SHELF protocol [18, 36], are the gold standard for formulating prior distributions, but require resources and expertise generally unavailable to a laboratory making routine measurements. It is for this everyday context that we aim for readily applied ways to specify prior distributions.

The noninformative prior distribution (referred to as NIP here) for σ^2 used in JCGM 101 [3] to derive the Bayesian analysis in section 2.1 is a limiting case of an ICS distribution in which the degrees of freedom parameter tends to zero.

As alternatives to the noninformative formulation, we propose two informative ICS distributions that can be assigned in practice based only on simple judgments, even when prior information is relatively weak. The SIP(v) and MIP(v) distributions have scale v and degrees of freedom 8 and 3, respectively.

B.1. The SIP distribution

Figure 2 (left) shows the SIP(v) distribution. As a prior distribution for a variance σ^2, it represents a judgment that σ^2 is most likely to be around the estimate v, and is highly likely to be in the interval $v/3$ to $3v$. Prior information about the measurement errors is more naturally expressed in terms of the standard deviation than the variance. The distribution of σ when $\sigma^2 \sim$ SIP(v) is shown in figure 2 (right).

Thus, the prior distribution $\sigma^2 \sim$ SIP(v) represents a belief that σ is most likely to be around its estimate of $\sqrt{v}$ and is highly likely to be within a factor $\sqrt{3} \approx 1.732$ of that value. It is judged almost certain to be in the interval $\sqrt{v}/2$ to $2\sqrt{v}$. For an organization carrying out regular testing with the same equipment, it is reasonable to suppose that there will be at least this level of knowledge of the standard deviation of measurement errors.

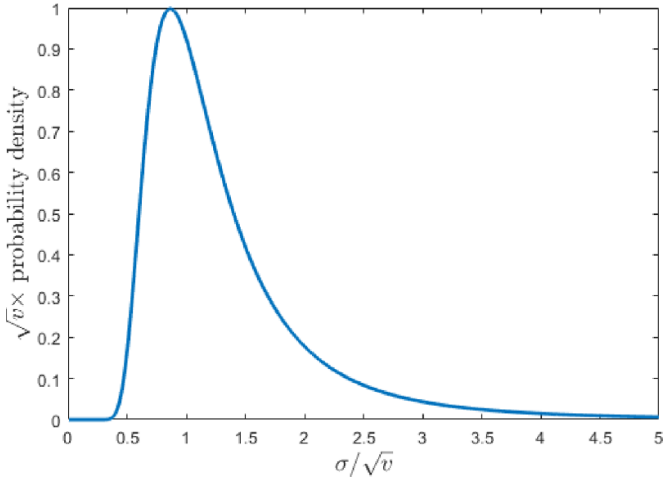

Figure 3. Density function of the square root of MIP(v).

B.2. The MIP distribution

The MIP(v) distribution is an alternative when there is less certainty about σ. Figure 3 shows the distribution of σ when $\sigma^2 \sim$ MIP(v). Thus, the prior distribution $\sigma^2 \sim$ MIP(v) represents a belief that σ is most likely to be around its estimate of $\sqrt{v}$ and is highly likely to be within the interval $\sqrt{v}/3$ to $3\sqrt{v}$.

B.3. One normal sample with ICS prior

We show here that ICS distributions are also a convenient choice because they allow a simple application of Bayes' theorem to combine the prior distribution with the information in the data. Formally, they are conjugate distributions [37] for the variance of a normal sample.

Consider the case of a single normal sample, as in section 2. Suppose that $\sigma^2 \sim vd\chi_d^{-2}$, and we assign a noninformative uniform prior to μ. Standard Bayesian analysis

[35] yields a posterior distribution with the following features.

- $\sigma^2 \sim v^* d^* \chi_{d^*}^{-2}$, where

$$d^* = d + n - 1, \qquad v^* = \frac{vd + (n-1)s^2}{d^*}.$$

  Thus, the $n-1$ degrees of freedom in the sample is augmented by the d degrees of freedom in the prior distribution. The posterior scale parameter v^* is a weighted average of the prior scale parameter v and the sample variance s^2, with weights proportional to their respective degrees of freedom.
- μ has a scaled and shifted Student's t distribution with mean $\bar{x}$ and scale parameter v^*/n. Its variance is $v^* d^*/[n(d^*-2)]$. For $d^* > 2$ (as it is with both the MIP and SIP prior distributions), the variance exists, even for a sample of size 1.

Therefore, a metrologist using an ICS prior distribution such as MIP(v) or SIP(v) will assign the median estimate $\bar{x}$, just as in the original GUM analysis, with characteristic uncertainty

$$c(\mu) = \frac{k(d^*)}{2}\sqrt{\frac{v^*}{n}}$$

as in expression (1). Notice that when $d = 0$ the value of v is irrelevant: it has no effect on the posterior distribution of either μ or σ^2. This is why we omit v when designating the noninformative prior distribution as NIP.

B.4. Alternative forms of algorithm 1

Section 2.4 presents the basic algorithm 1 for Type A evaluation with reporting of the measurement result in the form of (a) the median and characteristic uncertainty, and (b) the probability distribution of the measurand. This is the form advocated by [28] and described as 'meaningful expression of uncertainty in measurement' (MUM).

Although we strongly prefer the MUM form of reporting, we recognize that a metrologist may nevertheless wish to express uncertainty using a standard deviation. As noted in section 2.3, there are different definitions of standard uncertainty, with different interpretations, according to the frequentist and Bayesian statistical paradigms. The use of an informative prior distribution, whether MIP or SIP, implies adoption of the Bayesian paradigm and accordingly it would be natural to report the Bayesian standard uncertainty (BSU), which is the standard deviation of the posterior distribution.

Some metrologists are uncomfortable with the BSU, at least in the case of a NIP distribution, preferring the frequentist standard uncertainty $s/\sqrt{n}$ of the GUM. There is no genuinely frequentist standard uncertainty available when an informative prior distribution has been used, but an obvious analogue would by $\sqrt{v^*/n}$, which we refer to as a hybrid standard uncertainty (HSU).

Algorithm 1A. Type A evaluation from a normal sample with MIP or SIP prior information and reporting Bayesian (BSU) or hybrid (HSU) standard uncertainty.

Begin: Follow algorithm 1 until step 6 of the *Compute* section:
6. Form standard uncertainty
(BSU) $u(\mu) = \sqrt{d^*/(d^*-2)}\sqrt{v^*/n}$)
(HSU) $u(\mu) = \sqrt{v^*/n}$

Report:
(a) The measured value (mean) of μ is $\bar{x}$ with standard uncertainty $u(\mu)$
(b) (BSU) The quantity μ has a t distribution with mean $\bar{x}$, scale parameter v^*/n and degrees of freedom d^*
(HSU) The degrees of freedom associated with the measurement is d^*

End

B.5. Predictive distribution

We now derive the predictive distribution for the ratio s^2/v in section 2.6. Conditional on σ^2, s^2/v has the scaled chi-square distribution

$$\frac{s^2}{v} \mid \sigma^2 \sim \frac{\sigma^2}{v}(n-1)^{-1}\chi^2_{n-1}.$$

In general, if the prior distribution of σ^2 is ICS with degrees of freedom d and scale parameter v, then

$$\frac{\sigma^2}{v} \sim d\chi_d^{-2}.$$

Therefore the unconditional, that is, the prior predictive, distribution of s^2/v is

$$\frac{s^2}{v} \sim \frac{(n-1)^{-1}\chi^2_{n-1}}{d^{-1}\chi^2_d} = F_{n-1,d}.$$

Thus, in the case of a MIP(v) prior distribution, the predictive distribution is $F_{n-1,3}$, and in the case of a SIP(v) prior distribution it is $F_{n-1,8}$.

Appendix C. Computations in the examples

C.1. Computations for the single normal sample example

Section 2.5 presents results for a single normal sample, using NIP, MIP and SIP prior distributions for σ^2. In general, consider a sample of size n and a sample variance of s^2, and for convenience write $W = (n-1)s^2/\sigma^2$. The characteristic uncertainty for the NIP distribution (and for the frequentist analysis) is half the expanded uncertainty:

$$c(\mu) = \frac{1}{2}k(n-1)\frac{s}{\sqrt{n}} = \frac{1}{2}k(n-1)\sqrt{\frac{\sigma^2 W}{n(n-1)}}.$$

From expression (1), for the MIP(v) prior,

$$c(\mu) = \frac{1}{2}k(2+n)\sqrt{\frac{3v+\sigma^2 W}{n(2+n)}}, \tag{4}$$

and, for the SIP(v) prior,

$$c(\mu) = \frac{1}{2}k(7+n)\sqrt{\frac{8v+\sigma^2 W}{n(7+n)}}. \quad (5)$$

Since $W \sim \chi^2_{n-1}$, the expected values of these characteristic uncertainties can all be computed for given σ^2 and n by numerical integration or by a Monte Carlo computation.

The graphs of expected characteristic uncertainty for MIP and SIP priors in figure 1 (left) were computed for $n = 5$, $v = 1$ and each value of σ by Monte Carlo, sampling 10^6 values of W.

Section 2.5 also reports the median percentage reduction in characteristic uncertainty obtained by the MIP and SIP prior distributions, relative to the NIP distribution, and the probability that the reduction is positive. In each case, the calculations assume that σ^2 is drawn from the metrologist's prior distribution, MIP or SIP respectively. These were computed by Monte Carlo again, sampling 10^6 values of both W and σ^2. For each sampled pair, the characteristic uncertainties were calculated as given above and the percentage reduction using the MIP or SIP prior was computed. The median reduction was then computed as the median of the 10^6 percentage values and the probability of a reduction was computed as the proportion of percentage reductions that were positive.

For each prior distribution, the 95 % credible interval is $\bar{x} \pm 2c(\mu)$. For the NIP distribution, the coverage probability is 95 %, for all σ^2, but for the informative priors the coverage is a function of σ^2. To compute this coverage, we note that the credible interval contains the measurand value μ if

$$(\bar{x}-\mu)^2 \leqslant 4c^2(\mu),$$

and we now let $V = n(\bar{x}-\mu)^2/\sigma^2 \sim \chi^2_1$. The coverage probability can then be computed by a simple Monte Carlo computation. In the case of the MIP(v) prior, randomly draw a value V from the χ^2_1 distribution and a value W from the χ^2_{n-1} distribution, and then evaluate the condition

$$\frac{\sigma^2 V}{n} \leqslant \frac{[k(2+n)]^2(3v+\sigma^2 W)}{n(2+n)}$$

and therefore

$$fV - W \leqslant \frac{3v}{\sigma^2},$$

where

$$f = \frac{2+n}{[k(2+n)]^2}.$$

The coverage probability is then estimated by the proportion of times this condition is satisfied in a large number of simulated draws of (V, W). The corresponding condition for the SIP(v) prior is readily derived.

The graphs of coverage probability for MIP and SIP priors in figure 1 (right) were computed for $n = 5$ and $v = 1$ and each value of σ by Monte Carlo, sampling 10^6 values of V and W.

Note that the coverage condition does not depend on μ at all, and only depends on σ^2 and v through their quotient v/σ^2.

C.2. Computations for the multiple normal samples example

In the single burning item test example of section 3.2, the model (2) expresses the measurand κ in terms of six inputs. As shown in section 3.2, they are evaluated to have independent t distributions. If they were assigned MIP prior distributions the degrees of freedom for each input would be 6, while they would have 11 degrees of freedom if SIP prior distributions were assigned. For both prior distributions, the means of the t distributions are given in the second column of table 4. The scale parameters are given in the second or third column of table 5 in the case of MIP or SIP distributions respectively.

For example, in the case of the MIP prior, the input w_1 has a t distribution with 6 degrees of freedom, mean 7.00 and scale parameter 0.7657. Thus, its variance is $0.7657 \times 6/4 = 1.1486$ and its standard deviation is therefore 1.072.

Algorithm 2 was applied using the measurement model equation (2), the respective t distributions of the six input quantities and a Monte Carlo sample size of $M = 10^6$. If sampling from the t distributions arising from the MIP priors, the characteristic uncertainty was found to be $c(\kappa) = 0.052$, while if sampling from the t distributions arising from SIP priors then $c(\kappa) = 0.045$.

ORCID iDs

Anthony O'Hagan https://orcid.org/0000-0002-7994-0702
Maurice Cox https://orcid.org/0000-0002-6342-7840

References

[1] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2012 *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms* (Joint Committee for Guides in Metrology, JCGM 200)
[2] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2008 *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement* (Joint Committee for Guides in Metrology, JCGM 100)
[3] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2008 *Evaluation of Measurement Data—Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement"—Propagation of Distributions Using a Monte Carlo Method* (Joint Committee for Guides in Metrology, JCGM 101)
[4] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2011 *Evaluation of Measurement Data—Supplement 2 to the "Guide to the Expression of Uncertainty in Measurement"—Models with Any Number of Output Quantities* (Joint Committee for Guides in Metrology, JCGM 102)
[5] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2012 *Evaluation of Measurement Data—The Role of Measurement Uncertainty in Conformity Assessment* (Joint Committee for Guides in Metrology, JCGM 106)
[6] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML 2020 *Guide to the Expression of Uncertainty in Measurement—Part 6: Developing and Using Measurement Models* (Joint Committee for Guides in Metrology, GUM-6)
[7] Gleser L J 1998 Assessing uncertainty in measurement *Stat. Sci.* **13** 277–90

[8] Kacker R and Jones A 2003 On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent *Metrologia* **40** 235–48
[9] Lira I 2019 Type A evaluation of measurement uncertainty: frequentist or Bayesian? *2019 XXIX Int. Scientific Symp. "Metrology and Metrology Assurance" (MMA)* (IEEE)
[10] Forbes A B and Sousa J A 2011 The GUM, Bayesian inference and the observation and measurement equations *Measurement* **44** 1422–35
[11] Lira I and Grientschnig D 2010 Bayesian assessment of uncertainty in metrology: a tutorial *Metrologia* **47** R1–R14
[12] Possolo A and Meija J 2020 *Measurement Uncertainty: A Reintroduction* (Ottawa: National Research Council)
[13] Shirono K and Cox M 2019 Statistical reassessment of calibration and measurement capabilities based on key comparison results *Metrologia* **56** 045001
[14] Sené M, Gilmore I and Janssen J-T 2017 *Nature* vol 547 (London: Nature Publishing Group) 397–9
[15] Marschall M, Wübbeler G and Elster C 2022 Rejection sampling for Bayesian uncertainty evaluation using the Monte Carlo techniques of GUM-S1 *Metrologia* **59** 015004
[16] Demeyer S, Fischer N and Elster C 2021 Guidance on Bayesian uncertainty evaluation for a class of GUM measurement models *Metrologia* **58** 014001
[17] Carobbi C 2022 An informed type A evaluation of standard uncertainty valid for any sample size greater than or equal to 1 *Acta IMEKO* **11** 1–5
[18] O'Hagan A 2014 Eliciting and using expert knowledge in metrology *Metrologia* **51** S237–44
[19] Possolo A and Elster C 2014 Evaluating the uncertainty of input quantities in measurement models *Metrologia* **51** 339
[20] Stoudt S, Pintar A and Possolo A 2021 Uncertainty evaluations from small datasets *Metrologia* **58** 015014
[21] ISO 3534-2 2006 *Statistics—Vocabulary and Symbols—Part 1: Applied statistics* (International Organisation for Standardization)
[22] Tyler Estler W 1999 Measurement as inference: fundamental ideas *CIRP Ann.* **48** 611–31
[23] Lambkin D, Nortcliff S and White T 2004 The importance of precision in sampling sludges, biowastes and treated soils in a regulatory framework *TrAC Trends Anal. Chem.* **23** 704–15
[24] Papananias M, McLeay T E, Mahfouf M and Kadirkamanathan V 2019 A Bayesian framework to estimate part quality and associated uncertainties in multistage manufacturing *Comput. Ind.* **105** 35–47
[25] Cox M and Shirono K 2017 Informative Bayesian Type A uncertainty evaluation, especially applicable to a small number of observations *Metrologia* **54** 642–52
[26] van der Veen A M H 2018 Evaluating measurement uncertainty in fluid phase equilibrium calculations *Metrologia* **55** S60–S69
[27] ISO 28596 2022 *Sampling Procedures for Inspection by Attributes—Two-Stage Sampling Plans for Auditing and for Inspection Under Prior Information* (International Organisation for Standardization)
[28] Cox M and O'Hagan A 2022 Meaningful expression of uncertainty in measurement *Accreditation Qual. Assur.* **27** 19–37
[29] Huggett C 1980 Estimation of rate of heat release by means of oxygen consumption measurements *Fire Mater.* **4** 61–65
[30] McCaffrey B J and Heskestad G 1976 A robust bidirectional low-velocity probe for flame and fire application *Combust. Flame* **26** 125–7
[31] PD CEN/TR 16988 2016 Estimation of uncertainty in the single burning item test *Technical Report* (CEN)
[32] Bernard S, Marrelec G, Laugier P and Grimal Q 2015 Bayesian normal modes identification and estimation of elastic coefficients in resonant ultrasound spectroscopy *Inverse Problems* **31** 065010
[33] Yanof A 2001 Techniques and tools for photo metrology *Handbook of VLSI Microlithography* (Amsterdam: Elsevier) pp 382–471
[34] Lira I 2011 Monte Carlo evaluation of the uncertainty associated with the construction and use of a fitted curve *Measurement* **44** 2156–64
[35] O'Hagan A and Forster J 2004 *The Advanced Theory of Statistics: Bayesian Inference* vol 2B (London: Arnold)
[36] Oakley J E and O'Hagan A 2019 *SHELF: The Sheffield Elicitation Framework (Version 4)* (School of Mathematics and Statistics, University of Sheffield) (available at: http://tonyohagan.co.uk/shelf) (Accessed 10 March 2021)
[37] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* (Berlin: Springer)