AMT-2021-157 REFEREE REPORT

# Truth and Uncertainty at the Crossroads

Antonio Possolo, *NIST Fellow*
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

July 13, 2021

## Recommendation

The article's Abstract sums up the central claims accurately that the authors develop and substantiate in their narrative, including what I believe to be the correct conclusion that the "error" and "uncertainty" concepts are not fundamentally different, and may be regarded as alternative and complementary interpretations of the doubt about the true value of the measurand that remains after measurement.

However, my reading of the *Guide to the Expression of Uncertainty in Measurement* (GUM) [Joint Committee for Guides in Metrology (JCGM), 2008] suggests less polarizing views about this issue than the views that the authors of the article under review derive from the same Guide.

Their take on things brings to mind the acerbic discussion of essentially the same issues that took place in meetings of the ISO/TAG-4 Working Group 3, around 1986-87 [Collé, 1987a,b] [Schumacher, 1987].

The article should be published after it will have been shortened and more sharply focused to convey its message most effectively, and after improvements will have been made to deficient passages that are discussed under *Specific Comments*.

The *Technical Corrections* offer an assortment of suggestions concerning English usage that the authors should consider.

## General Comments

Acknowledgment should be made of an understanding of the relation between measurement uncertainty and measurement error that predates the GUM, and that the authors of the article under review likely will find agreeable:

> 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 [Eisenhart and Collé, 1980].

Churchill Eisenhart was my most illustrious predecessor at NIST, and Ronald Collé, a distinguished and esteemed NIST colleague, served as convener of the working group (ISO-TAG-4/WG3) that laid the groundwork for the creation of the GUM [Collé and Karp, 1987].

A discussion whose tenor places the "error approach" on Mars and the "uncertainty approach" on Venus sounds more like the discussions that inflamed the metrological community thirty-five years ago, than a useful discussion that we can engage in today with the benefit of the experience accumulated in these many intervening years [Eisenhart and Collé, 1980] [Collé, 1987a] [Colclough, 1987] [Schumacher, 1987].

The viewpoint that the authors of the article under review wish to convey, can be conveyed quite simply also by means of an allegory: measurement errors are the "carriers" of measurement uncertainty, in a sense analogous to how photons are the "carriers" of light waves and, more generally, of the electromagnetic force.

Accepting such dualism between errors and uncertainty facilitates the scientific discourse without excluding individual or cultural preferences, and tones down the drama that has been unfolding in the literature and that, at times, this article also exacerbates unnecessarily.

The 26 pages of text of the article under review arguably are overkill to convey this simple, conciliatory message: what they do prompt is a review almost as long as the article itself, thus making this review much too long by any standard.

In fact, the key message of the article will be delivered more effectively, and the article will have greater impact, if the article is shortened and its arguments are streamlined.

The article's length can be reduced at least by deleting those portions that distract more than they add insight: for example, the digressions in section 2 and in subsections 5.2 and 5.3.

The authors may wish to extend their criticism to the *International Vocabulary*

*of Metrology* (VIM) [Joint Committee for Guides in Metrology, 2012], whose Introduction states:

> The change in the treatment of measurement uncertainty from an Error Approach (sometimes called Traditional Approach or True Value Approach) to an Uncertainty Approach necessitated reconsideration of some of the related concepts appearing in the second edition of the VIM.

The authors also seem to be unaware of the critical evaluation of the GUM that Gleser [1998] published shortly after the original, 1993 edition of the GUM was corrected and reprinted, in 1995 [BIPM et al., 1995]. References to suitable portions of this evaluation will add value to the article under review, and will also facilitate shortening it.

The article is very repetitive in the multiple instances where it rehashes the relations between the concepts of *error*, *uncertainty*, *true value*, and *Bayesianism*. Consolidating and refocusing the fragmentary discussion of these relations would make the article much easier to read and would enhance the cogency of its arguments. However, accomplishing this would involve a major rewrite.

In its Annex E (E.5.1) the GUM addresses the issue that is the main focus of the article under review, when it states that

> The focus of this Guide is on the measurement result and its evaluated uncertainty rather than on the unknowable quantities "true" value and error (see Annex D). By taking the operational views that the result of a measurement is simply the value attributed to the measurand and that the uncertainty of that result is a measure of the dispersion of the values that could reasonably be attributed to the measurand, this Guide in effect uncouples the often confusing connection between uncertainty and the unknowable quantities "true" value and error.

The authors of the article under review quite correctly point out that the uncertainty is neither a property of the measured value nor is it *about* the measured value. The uncertainty surrounds or clouds the true value, and qualifies the state of knowledge that the metrologist has of the true value.

To the extent that the target of measurement is the true value, the measurement error is meaningful even if not observable (however, it can be estimated in some cases, as discussed below in relation with Line 180).

The suggestion, made in the aforementioned E.5.1, that "uncertainty," "error," and "true value" should be uncoupled from one another seems at odds with what is actually done in the practice of measurement science. For example, in relation with certified reference materials, "NIST asserts that a certified value provides an estimate of the true value of a defined measurand" [Beauchamp et al., 2020, 1.2.4].

Therefore, the implied understanding of the scientists developing these materials is that the uncertainties reported in the corresponding certificates are informative about the relation between the measured value and the true value, the difference between the former and the latter being the measurement error.

## Specific Comments

The numbers in boldface refer to line numbers in the version of the preprint made available for discussion on June 29, 2021.

**002 + 245** Here and elsewhere throughout the article, "GUM8" should be replaced by "GUM" because the GUM and its existing and planned supplements are being rearranged and renumbered, and "GUM8" is already reserved to refer to something other than the current GUM. For similar reasons, "GUM09" is likely to be misinterpreted, and should not be used: in fact, it is not needed at all because the authors use this acronym only in the very same line (245) where they introduce it.

**010** *the term 'error' was used, with some caveats, for designating a statistical estimate of the expected difference between the measured and the true value of a measurand*

The traditional and still customary meaning of "error" in statistical models is of a non-observable difference between the observed and the (generally also non-observable) true value of a quantity [Davison, 2008, Example 1.1].

For example, in the relationship $m = \mu + \varepsilon$ between a measured value, $m$, and the true value, $\mu$, of the mass of a massive entity, $\varepsilon$ is the error.

The error is generally neither known nor observable, but in many situations it can be estimated, with $\widehat{\varepsilon}$ being commonly used to denote the estimate (refer to the discussion of Line 180).

In the discussion of Lines 518 + 738 below, it will become clear how useful the explicit consideration of error can be, by allowing one conceptually to separate contributions made by different sources of uncertainty.

**015** *stipulated a new terminology, where the term 'measurement uncertainty' is used in situations where one would have said 'measurement error'*

The word "error" occurs 131 times throughout the GUM, and not always deprecatingly. For example, in its 2.2.4, the GUM acknowledges that "The definition of uncertainty of measurement [...] is not inconsistent with other concepts of uncertainty of measurement, such as a measure of the possible error in the estimated value of the measurand."

**025** *the error statisticians and the uncertainty statisticians*

This classification of statisticians into these two classes is an invention of the authors that is more reflective of their imagination than of reality. In fact, the principal participants in the debates that took place in and around the aforementioned ISO/TAG-4/WG-3 were not statisticians.

Furthermore, the issue of "systematic" versus "random" errors (which we will discuss below, in relation with Line 597) may have been even more divisive than the issue of "error" versus "uncertainty."

Therefore, I urge the authors to devise a different way of characterizing the two camps they are alluding to here. A reference to Mayo and Spanos [2011] would be appropriate.

**046** *according to Bayesian statistics (Bayes, 1763) the measured value cannot always be interpreted as the most probable value of the measurand*

Since one does not need to invoke Bayesian statistics to reach the same conclusion [Possolo and Iyer, 2017, Page 011301-12], this remark is spurious.

**071** *Recapitulation of the concept of indirect measurements*

I believe that this long foray into inverse problems adds nothing of value to the discussion, hence suggest that section 2 be deleted. The discussion in subsection 6.3, *The causal arrow*, can easily be reformulated, and in the process also shortened, to drive the same points across — refer to specific suggestions below, for Line 618.

**119** *ancient researchers realized that measurement results always have errors*

It is all a matter of perspective, of course, but I am of the opinion that it is unfair to call Gauss or Legendre "ancient." In the context of European history, the word is typically reserved for the period ending with the fall of the Western Roman Empire (around 500 CE).

In addition, and in particular concerning Gauss, with whose works I am more familiar than with Legendre's, I can only say that it is difficult for me to imagine

a person of more luminous modernity, or with a better sense for what is relevant in scientific practice (of his time or contemporary), than Gauss.

**149** *In the case of 'error', its statistical estimate is mostly understood to be a quadratic estimate and thus does not carry any information about the sign of the error.*

The authors may like to replace this awkward sentence with something along the following lines: "In most cases, errors are not estimated individually. Instead, their typical size is summarized by the square root of their mean squared value, or by the median of their absolute value. Such summaries do not preserve information about the signs of any individual errors."

**154** *the term 'error' has commonly been used to signify a statistical estimate of the size of the difference between the measured and the true value of the measurand*

This is repetitive of the material around Line 10 that was discussed above. In both instances, the authors are unnecessarily turning something simple into something complicated.

One thing is the error $\varepsilon$ in the example discussed above, $m = \mu + \varepsilon$. Another thing is how this error may be characterized or quantified.

For example, the possible errors may be characterized by the probability distribution of $\varepsilon$, like when one says: the signal was corrupted by white noise with mean 0 and standard deviation $\sigma$.

The sizes of possible errors may be summarized by the mean squared error (MSE) of the estimator of the measurand, which captures the difference between expected value of the estimator and the true value of the measurand, as well as dispersion around that expected value. Other summaries include the standard deviation of the error distribution, or the now outmoded *probable error*.

The error may also be characterized indirectly, by an expression of the uncertainty surrounding the quantity of interest. For example, Yoshino et al. [1988] reported the measurement result for the absorption cross-section of ozone at 253.65 nm as $1145^{+7.1}_{-14.4} \times 10^{-20}\text{cm}^2/\text{molecule}$, which says that the measurement error has an asymmetric distribution.

**180** *Since the true value is not known, the actual difference between the measured or estimated value and the true value of the measurand cannot be calculated.*

The authors quite correctly point out that this argument lacks cogency. In fact, more can be said further to dismiss this claim as being no more than a myth.

Consider the simplest of cases of statistical estimation, where one has replicated

determinations of the same quantity, $r_1, \ldots, r_m$, which are then combined to obtain an estimate $t = T(r_1, \ldots, r_m)$ of a quantity $\tau$. The estimate $t$ could be as simple as the average or the median of the replicates, or it could be their coefficient of variation (standard deviation divided by the average).

It is then possible, using the statistical jackknife [Mosteller and Tukey, 1977, Chapter 8] or the statistical bootstrap [Efron and Tibshirani, 1993], to estimate not only the standard deviation of $t$ (based on this single set of replicates $\{r_i\}$), but also both the sign and the magnitude of the error $t - \tau$.

**200 + 364** *our reading is that an error distribution is understood as a distribution whose spread is the estimated statistical error and whose expectation value is the true value, while an uncertainty distribution is understood as a distribution whose spread is the estimated uncertainty and whose expectation value is the measured or estimated value / The error distribution must not be conceived as a probability density distribution of a value to be the true value*

In the simple model for a measured mass, $m = \mu + \varepsilon$, the "error distribution" generally refers to the probability distribution of $\varepsilon$, hence the expected value of the "error distribution" will not be $\mu$, which denotes the true value of the measurand. Instead, this expected value will be the *bias*, which is the persistent offset of $m$ from $\mu$. (Refer to the discussion of Line 597, where I explain why I prefer "persistent" to "systematic," and "volatile" to "random.")

Neither "error distribution" nor "uncertainty distribution" are mentioned in the GUM. While the GUM offers considerable guidance about the assignment of distributions to input quantities, $\{x_j\}$, in its first 69 pages (out of a total of 120) all that it provides about the probability distribution of the output quantity, $y$, is an approximation to its standard deviation, in Equations (10) and (13).

Furthermore, the GUM seems to be more concerned with evaluating $u(y)$ than with estimating the measurand optimally, because the "substitution" estimate of the measurand, which is obtained by substituting the $\{x_j\}$ by their best estimates in $y = f(x_1, \ldots, x_n)$, generally will not yield the best estimate of the measurand in the sense of minimizing mean squared error, mean absolute error, or other similar criteria [Possolo and Iyer, 2017, Page 011301-12].

In the course of those initial 69 pages, the GUM touches upon the topic of the distribution of $y$ tangentially — for example when it discusses expanded uncertainty, coverage factors, and coverage interval —, but only in its Annex G (beginning on Page 70) does the GUM venture into a discussion of how to characterize the probability distribution of $y$.

Annex G invokes the Central Limit Theorem based on a first-order Taylor approximation of the measurement function $f$ in $y = f(x_1, \ldots, x_n)$, to claim that

$y$'s distribution may be taken as being approximately Gaussian. This argument can, on occasion, be spectacularly inaccurate [Possolo, 2015, Example E11].

Since $u(y)$ typically is based on finitely many degrees of freedom, the GUM argues (using a slightly different notation) that $(y - \eta)/u(y)$, where $\eta$ denotes $y$'s true value, should have a Student's $t$ distribution approximately, wherefrom coverage intervals then issue readily, thus achieving the goal, stated in its clause 0.5, of providing "an interval about the measurement result that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the quantity subject to measurement."

The meaning of this distribution that the GUM, by hook or by crook assigns to $y$, and, even more importantly, the meaning of the distributions derived for $y$ by application of the Monte Carlo method, and of the coverage intervals based on them, should be the more appropriate and productive targets for critical review, similarly to what Stoudt et al. [2021] have done.

**213** *Frequentist statistics, we understand, is a concept where the term 'probability' is defined via the limit of frequencies for a sample size approaching infinity. This definition is untenable because it involves a circularity*

The authors oversimplify and are unacceptably dismissive.

If the Frequentist interpretation of probability were this "obviously" defective, then none of John Venn, Richard von Mises, Andrey Nikolaevich Kolmogorov, Jerzy Neyman or Jack Kiefer — all intellectual giants in their own right — would have embraced it.

I suggest that the authors avoid embarrassment by considering the excellent overview of the interpretations of probability compiled by Hájek [2007].

**265** *The values the rational agent believes to be true are sufficient in this case, because the error distribution does not tell us anything about the truth anyway but only about the agent's believe of what truth is.*

The simplest measurement error model mentioned above, $m = \mu + \varepsilon$, is meaningful under essentially all paradigms of statistical inference.

In neither the classical (Frequentist) nor in the Bayesian approaches does the probability distribution of $\varepsilon$ convey any information about $\mu$, other than in special cases: for example, when the variance of $\varepsilon$ depends on $\mu$.

Both approaches involve assigning a probability distribution to $\varepsilon$, which then determines the likelihood function. The Bayesian approach involves also assignments of probability distributions to $\mu$ and to any parameters in the distribution of $\varepsilon$ whose values are unknown.

The marginal distribution of $m$ typically will differ in the classical and Bayesian approaches even when the same choice is made for the distribution of $\varepsilon$.

**317** *Monte Carlo uncertainty estimation, however, is in its heart a frequentist method, because it estimates the uncertainty from the frequency distribution of the Monte Carlo samples.*

The authors are quite wrong on this one.

Of course, the extent of how wrong depends on what they mean by "Monte Carlo uncertainty estimation." I assume that they mean it in the sense and context in which it was introduced into uncertainty analysis by Morgan and Henrion [1992], subsequently having been incorporated into the GUM Supplement 1 [Joint Committee for Guides in Metrology, 2008].

In such sense and context, the Monte Carlo method is purely mathematical, and non-denominational (neither Frequentist nor Bayesian), and solves the following problem: given a random vector $\boldsymbol{X}$ whose probability distribution has been fully specified, and a real-valued, measurable function $f$ defined on the range of $\boldsymbol{X}$, determine the probability distribution of $Y = f(\boldsymbol{X})$.

The Monte Carlo method solves this problem using numerical methods and sampling driven by pseudo-random numbers. It solves it in the sense that it can produce the value of $\Pr(Y \in B)$ to within any specified accuracy, for any measurable subset $B$ in the range of $Y$.

The fact that its accuracy is guaranteed by the Law of Large Numbers does not make it Frequentist because the Law of Large Numbers is neither Frequentist nor Bayesian. The Law of Large Numbers is a mathematical result about sums of random variables based on Kolmogorov's axioms for probability measures [Kolmogorov, 1933].

If the authors' views on the Monte Carlo method were correct, then Markov Chain Monte Carlo sampling, which is the workhorse of contemporary Bayesian inference, would be "in its heart a frequentist method" too!

**318** *it is astonishing why GUM08, if representing a Bayesian concept, does not in the first place require to apply the Bayes theorem*

The authors should reference Gleser [1998] who points out the mixed-bag of viewpoints coexisting in the GUM. Clearly the authors are well entitled to feel astonishment at the GUM not using Bayes rule at all, especially considering the whirlwind of claims about the GUM and its Supplements being Bayesian.

However, in fairness to the GUM, such whirlwind has been more of an afterthought than a consequence of the GUM itself. First, the word "Bayes" is nowhere to be found in the GUM, and the word "Bayesian" occurs exactly once:

in the title of reference [14], on Page 115.

Only in Annex E (E.3.5) does the GUM venture into this controversial territory when it says "In contrast to this frequency-based point of view of probability, an equally valid viewpoint is that probability is a measure of the degree of belief that an event will occur." And then it adds: "Recommendation INC-1 (1980) upon which this Guide rests implicitly adopts such a viewpoint of probability."

The expression "degree of belief" occurs exactly once in the main body of the GUM (3.3.5), where it says:

> Thus a Type A standard uncertainty is obtained from a probability density function (C.2.5) derived from an observed frequency distribution (C.2.18), while a Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [often called subjective probability (C.2.1)]. Both approaches employ recognized interpretations of probability.

The same expression occurs in Annex C, and again in Annex E, where E.3.6 comes the closest to advocacy by enumerating "three distinct advantages to adopting an interpretation of probability based on degree of belief."

Therefore, and on the whole, the GUM is far more discreetly or ambiguously Bayesian than it has more recently been heralded to be (surprisingly, mostly by "born again," self-declared Bayesians).

The GUM's alleged Bayesianism in fact reduces to (i) entertaining (subjective) probability distributions for input quantities that are elicited from experts, and (ii) regarding the probability distribution of the measurand as quantification of degrees of belief about the true value of the measurand, even though it is not a Bayesian posterior distribution [Gleser, 1998, 2.2].

**343** *This suggests that the uncertainty is an attribute of the true value while the error is associated with a measurement or an estimate. Because of the measurement error there is an uncertainty as to what the true value is. The uncertainty thus describes the degree of ignorance about the true value while the estimated error describes to which degree the measurement is thought to deviate from the true value*

The authors are quite right. Please consider the following rewrite, which, although allegorical, I believe further enhances the expression of the authors' sentiment — also compare with Possolo [2015, Note 3.2, Page 16]:

> This suggests that measurement uncertainty surrounds the true value

> of the measurand like a fog that obfuscates it, while measurement error is both the source of that fog and part and parcel of the measured value. Measurement uncertainty thus describes the doubt about the true value of the measurand, while measurement error quantifies the extent to which the measured value deviates from the true value.

**379** *The weight of Thomas Bayes or the body height of David Hume at a certain time are well-defined quantities although we have no chance to measure them today*

I suggest that, for the sake of propriety and good taste, the authors abstain from referring to properties of the bodies of Thomas Bayes and David Hume, refined and excellent gentlemen both, long deceased, and use instead properties of other notable material entities that are no longer amenable to measurement, like the Colossus of Rhodes or the Lighthouse of Alexandria.

**411** *5.2 Likelihood, probability, and the base rate fallacy*

I believe that this subsection is a digression from the main topic that would best be deleted. A shorter, better focused article will have greater impact than one with multiple digressions that are largely off-topic.

**481** *5.3 Nonlinearity issues*

The same suggestion as for subsection 5.2, for the same reasons.

**518 + 738** *5.4 Incompleteness of the error budget*

This is an important issue that the authors should address in greater generality than in the context of inverse problems. The following example captures the key issues clearly and simply. The authors allude to the same ideas in Line 738.

The values measured in inter-laboratory studies are often modeled as $m_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \ldots, n$, where $\mu$ denotes the true value of the quantity of interest, and the $\{\lambda_j\}$ and the $\{\varepsilon_j\}$ are errors of different kinds: the former express laboratory effects [Toman and Possolo, 2009a,b, 2010], which in many cases will be persistent effects attributable to differences between measurement methods or between forms of calibration; the latter are laboratory-specific measurement errors quantified in the uncertainties reported by the participants.

The reality of the $\{\lambda_j\}$ (that is, that they cannot all be zero) becomes apparent only when the measurement results are put on the table and inter-compared. If the measured values are significantly more dispersed than the associated, reported uncertainties intimate that they should be, then this is an indication

that there is some *dark uncertainty* [Thompson and Ellison, 2011] afoot that was not captured in the individual uncertainty budgets.

This dark uncertainty is "carried" (in the sense in which this term was used in the *General Comments*) by the $\{\lambda_j\}$. Refer to Koepke et al. [2017] and to Possolo et al. [2021] for more extended discussions of this concept.

**548** *We have mentioned above that the uncertainty concept depends on the acceptance of the subjective probability in the sense of degree of rational belief. Without that, an error budget including systematic effects would make no sense because systematic effects cannot easily be conceived as probabilistic in a frequentist sense; that is to say, the resulting error cannot be conceived as a random variable in a frequentist sense.*

These statements are inaccurate.

First, the uncertainty concept may be contingent on a Bayesian perspective, but this perspective need not be subjective: it can be a so-called "objective Bayesian" perspective, which Jeffreys [1946], Bernardo [1979], and Berger [2006], among others, have favored.

Second, the main difficulty facing a Frequentist approach to the characterization of measurement uncertainty concerns what the GUM calls Type B evaluations of uncertainty components, not the recognition of the contributions that persistent ("systematic") effects make to said uncertainty.

In fact, the contributions from some persistent effects can be evaluated by Type A methods (refer to the comments above for line 180), and the contributions from some volatile ("random") effects can be evaluated by Type B methods (for example, the imprecision of a balance that a laboratory technician has great familiarity with).

**597** *Von Clarmann et al. (2020) explicitly demand that error estimates be classified as random or systematic [. . . ] In summary, the denial of the importance of distinguishing between random errors and systematic errors does not provide proper guidance, and altogether is a strong misjudgment.*

The word "demand" appears to be too strong a descriptor of what von Clarmann et al. [2020] actually did, which was to "formulate *recommendations* with respect to the evaluation and reporting of random errors, systematic errors, and further diagnostic data," where the emphasis on "recommendations" is mine.

We need to discuss two separate issues regarding this point: the first concerns the choice of terms ("systematic" and "random"); the second concerns whether and when to bundle them all into a single expression of uncertainty.

Concerning the first issue — the choice of terms:

My dislike of terms like "systematic" and "random" is that they are metaphysical: they speak to the nature of the errors, which is often elusive and may be shifting.

For example, von Clarmann et al. [2020, R3, Page 4420] recognize that "depending on the application of the data, the same type of error can act as random or systematic error," and many other authors have acknowledged the same. "Random," in particular, is a thorny concept, whose definition seems to be far from settled [Landsman, 2020] [Eagle, 2016] [Bennett, 2011] [Gács, 2005].

For these reasons, I recommend descriptive qualifiers instead, for example *persistent* (instead of "systematic") and *volatile* (instead of "random"). They are less committal and afford greater flexibility, in particular to address cases where a volatile error becomes persistent, or vice versa.

Writing almost thirty-five years ago, Collé [1987a], summarized the two approaches to measurement uncertainty that were then dominant as follows:

> The "*classical*" approach is based on a central distinction between so-called random and systematic uncertainties. The uncertainties are presumably classified by the underlying physical error type [...] and the approach demands that the different uncertainty types be combined by different methods. Causing even further confusion, the uncertainties in these classical treatments are said to depend on one's "perspective" and hey possess chameleon-like properties, and may change from one type to another.
>
> In contrast, the "*romantic*" approach dispenses with the underlying error distinction, and classifies the uncertainties only on the basis of how the uncertainty estimates were made. All uncertainty components in this approach can be combined by the same general propagation formulae. The romantic approach underlies the BIPM/CIPM Recommendation.

Concerning the second issue — the bundling of contributions from all sources of uncertainty:

While agreeing with the *romantic* approach in principle, I believe that it is advisable to consider how uncertainty evaluations will be used, before deciding whether to combine contributions from all sources of uncertainty into a single evaluation, or not. This is a more nuanced, less extreme approach than either of the two approaches aforementioned.

Consider an inter-laboratory study where several laboratories measure the same quantity independently of one another, or a meta-analysis of results from preexisting studies that were carried out and published independently of one another.

Suppose that the purpose is to blend the corresponding estimates into a consensus value: for example, as was done for the ozone absorption cross-section at 253.65 nm [Hodges et al., 2019].

Typically, the consensus value will be some form of weighted average. Therefore, the errors behind the uncertainties reported by the participants will "average out" in the process to some extent. This may be fine, or it may be inappropriate.

Such "averaging out" will be fine if laboratory-specific persistent errors lead to estimates that are high for some laboratories and low for other laboratories, with the true value lying somewhere in the middle.

But such "averaging out" will be inappropriate if a common bias, unbeknownst to all, affects all results similarly. Reporting separately the evaluation of the contributions made by persistent effects, and by volatile effects, as is commonly done in astrophysics and in particle physics, will then be an appropriate, prudent way to report uncertainty intended for use by a downstream user.

The need for such discretion, and the role that considerations of fitness-for-purpose of uncertainty evaluations should play in deciding what to do and when, is mentioned already in the pre-GUM literature [Ku, 1980].

**618** *6.3 The causal arrow — [. . . ] We think that it is essential to appreciate the inverse nature of the problem, and this is much easier if the measurement equation describes the forward problem and thus does not suggest an unambiguous determination of the measurand from the measured quantity.*

The measurement model in the GUM is only one of many kinds of measurement models to which the principles for uncertainty evaluation that are enunciated in the GUM apply. The GUM-6 [Joint Committee for Guides in Metrology, 2020], published recently, describes several other kinds of measurement models, including statistical measurement models.

Rodgers [2000, 2.3.2] explains how Bayesian statistical models can be used in general to solve inverse problems, and Ganesan et al. [2014] describe an application of hierarchical Bayesian methods to atmospheric trace gas inversions. The Bayesian approach can be fruitful in such settings because the prior distribution acts as a regularization prescription.

Possolo [2015] gives examples of measurements involving models that are quite different from the conventional measurement model in the GUM. In particular, Examples E7 (Thermistor Calibration), E17 (Gas Analysis), and E32 (Load Cell Calibration) concern calibrations that are structurally similar to the thermometer example that the authors mention in Line 109.

Using $x_1, \ldots, x_n$ and $y$ with the same roles that the GUM gives them, a statistical forward model can be formulated simply by saying $x_1, \ldots, x_n \sim L_y$, which is shorthand for "the joint probability distribution of (the random variables whose realized values are) the observable inputs $x_1, \ldots, x_n$ has $y$ as a parameter and likelihood function $L_y$.

A Bayesian formulation will then add $y \sim P$, where $P$ is the prior distribution of $y$, and application of Bayes's rule produces a solution for the inverse problem in the form of the posterior distribution, $Q$, of $y \sim Q_{x_1,\ldots,x_n}$. Compare this formulation with the treatment of calibration via conventional regularization in Hagwood [1992].

**662** *paradoxes shatter the bedrocks of Bayesian philosophy, namely the likelihood principle that says that all relevant evidence about an unknown quantity obtained from an experiment is contained in the likelihood. Others accept the theoretical validity of the Bayes theorem but challenge its applicability in real life because of the unknown and unknowable prior probabilities.*

The paradoxes alluded to often relate more to the adoption of so-called "non-informative" prior distributions than to the acceptance of the likelihood principle, as Cox [2006] points out, in a contribution referenced by White [2016].

All theories of inference have given rise to paradoxes, and nevertheless most often they produce valid and practically useful inferences. Regarding the likelihood principle in particular, at least one well-known "paradox" has been dismissed as a false alarm [Goldstein and Howard, 1991].

In any case, White [2016] does not come even close to suggesting that such paradoxes "shatter the bedrocks of Bayesian philosophy," in particular as applied in measurement science. I know for a fact that Rod White does not object to the use of Bayesian methods when these are warranted and there is genuine prior information that should be taken into account.

The objection, which is also raised by Bayesians [O'Hagan, 2006], is to the systematic reliance on "non-informative" prior distributions just for the sake of going through the motions of the Bayesian machinery or to pay lip service to scientific objectivity.

The Bayesian approach to problems of statistical inference is a choice among many that can be made, similarly to how some people choose to drink lemonade and others bourbon. Different approaches to statistical inference (be they frequentist, fiducial, or Bayesian) all can claim notable successes in solving problems of practical importance.

Bayesian methods, in particular, can boast a long and varied roster of accom-

plishments that prove beyond reasonable doubt that they are applicable in real life, and that they can be used to solve important practical problems, and that often they do so better than non-Bayesian alternatives [O'Hagan, 2008].

A particularly striking, recent accomplishment of Bayesian methods concerns the use of measurements of $\Delta^{14}CO_2$, in conjuction with atmospheric transport models, to demonstrate that several bottom-up approaches to the estimation of national inventories likely underestimate U.S. fossil fuel $CO_2$ emissions [Basu et al., 2020].

This study, which is based on methodological advances published in this very journal [Basu et al., 2016], includes rigorous, model-based uncertainty evaluations, and also serves to show that the GUM and its supplements have much catching-up to do if they will ever come to play a role in addressing momentous issues like the measurement of greenhouse gas emissions.

The suggestion that Bayesian methods are questionable because prior distributions are "unknown and unknowable" reveals a misconception about prior distributions: they are meant to encapsulate the knowledge that someone has about the quantity of interest, prior to performing an experiment that generates fresh information about it. Therefore, proper, informative, subjective prior distributions are known to who formulates them, by construction.

Of course, the Bayesian can be much mistaken and construct a prior distribution that reflects an erroneous conception of reality, in which case the "knowledge" that the prior encapsulates is false knowledge and its use will lead the inference astray. However, Bayesian methods cannot be blamed for delusions any more than Newton's laws can be blamed for accidental falls.

**674** *the Bayesian philosophy relies on a couple of unwarranted assumptions, e.g., the likelihood principle and the indifference principle.*

The authors convey a wrong impression on both counts.

Adherence to the likelihood principle is a choice that, in most applications, turns out to be a better choice than most alternatives. Still, it is only a choice, among many that can be made. Making such choice is necessary but not sufficient to be Bayesian. Many statisticians, physicists, chemists, and biologists adhering to the likelihood principle are not Bayesian.

Neither is adopting an indifference principle (or, more generally, using an allegedly non-informative prior distribution) necessary to qualify as being Bayesian. In fact, quite the contrary is true: reliance on proper, informative, and suitably elicited subjective prior distributions, are the hallmarks of genuine Bayesian practice. But this, too, is only a choice [Robert, 2007].

## Technical Corrections

**025** Replace *comes down to the question if and how* with "comes down to the question of whether, and if so how"

**056** Replace *Second we assess to which degree* with "Second, we assess the degree to which"

**068** Replace *we conclude to which degree* with "we conclude the degree to which"

**128** Replace *A rich methodical toolbox* with "A rich methodological toolbox"

**180** Replace *This argument is often used to dispraise* with "This argument is often used to disparage"

**267** Replace *agent's believe* with "agent's belief"

**364** Replace *Quantities of which the value cannot determined* with "Quantities whose values cannot be determined." This suggestion deliberately ignores the antiquated invective against using the possessive *whose* for inanimate objects, consistently with the recommendation in O'Conner [2019, Page 243].

**373** Replace *Others have been formulated by us, serving, as arguments of the Devil's advocate, as working hypotheses in order to moot the error and uncertainty concepts in the context of indirect measurements* with "We have formulated others as Devil's advocates, which are intended to serve as working hypotheses to MOOT the error and uncertainty concepts in the context of indirect measurements," except that "moot" needs to be replaced by a word that is suitable for this passage: maybe "merge" or "reconcile", depending on what the authors wish to express.

**413** Replace *measurements are not in the focus* with either "measurements are not in focus" or "measurements are not the focus," depending on what the authors wish to say exactly.

**420** Replace *the probability that a person suffering fever to have Covid-19 is 50%* with "the probability is 50 % that a person with fever has COVID-19"

**429** Replace *distribution which is missing* with "distribution that is missing"

**470** Replace *the aggregarion of random uncertainties* with "the aggregation of random uncertainties"

**612** Replace *strong misjudgement* with "strong misjudgment" (unless the British spelling be preferred)

**743** The sentence that includes *traditional error analysis can connotate a statistical quantity* is unclear, and should be rewritten, taking into account the fact that the verb *connotate* is obsolete and has been replaced by *connote*. However, a term more generally familiar would be preferable, like *suggest*, possibly.

## Acknowledgments

I thank my NIST colleagues Blaza Toman and David Newton (both from the Statistical Engineering Division), for the suggestions for improvement that they offered in relation with a draft of this contribution.

I am particularly indebted to Ronald Collé (Radioactivity Group of the Radiation Physics Division of NIST), who generously shared many of his recollections and records of the early stages of the development of the guidance that ultimately found its way into the GUM.

## References

S. Basu, J. B. Miller, and S. Lehman. Separation of biospheric and fossil fuel fluxes of $CO_2$ by atmospheric inversion of $CO_2$ and $^{14}CO_2$ measurements: Observation System Simulations. *Atmospheric Chemistry and Physics*, 16(9): 5665–5683, 2016. doi: 10.5194/acp-16-5665-2016.

S. Basu, S. J. Lehman, J. B. Miller, A. E. Andrews, C. Sweeney, K. R. Gurney, X. Xu, J. Southon, and P. P. Tans. Estimating US fossil fuel $CO_2$ emissions from measurements of $^{14}C$ in atmospheric $CO_2$. *Proceedings of the National Academy of Sciences*, 117(24):13300–13307, 2020. doi: 10.1073/pnas.1919032117.

C. R. Beauchamp, J. E. Camara, J. Carney, S. J. Choquette, K. D. Cole, P. C. DeRose, D. L. Duewer, M. S. Epstein, M. C. Kline, K. A. Lippa, E. Lucon, K. W. Phinney, A. Possolo, K. E. Sharpless, J. R. Sieber, B. Toman, M. R. Winchester, and D. Windover. *Metrological Tools for the Reference Materials and Reference Instruments of the NIST Materials Measurement Laboratory*. NIST Special Publication 260-136 (2020 Edition). National Institute of Standards and Technology, Gaithersburg, MD, 2020. doi: 10.6028/NIST.SP.260-136-2020.

D. Bennett. Defining randomness. In P. S. Bandyopadhyay and M. R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of*

*Science*, pages 633–639. North-Holland, Amsterdam, 2011. ISBN 978-0-444-51862-0. doi: 10.1016/B978-0-444-51862-0.50020-4.

J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3): 385–402, 2006. doi: 10.1214/06-BA115.

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41:113–128, 1979. doi: 10.2307/2985028.

BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML. *Guide to the expression of uncertainty in measurement (GUM)*. International Organization for Standardization (ISO), Geneva, Switzerland, 1995. ISBN 92-67-10188-9. Corrected and Reprinted.

A. R. Colclough. Two theories of experimental error. *Journal of Research of the National Bureau of Standards*, 92(3):167–185, May-June 1987. doi: 10.6028/jres.092.016.

R. Collé. Minutes of the meeting on measurement uncertainties. *NCSL Newsletter*, 27(4):52–55, October 1987a.

R. Collé. Report of the second meeting of the iso/tag-4/wg-3 working group "uncertainties". *NCSL Newsletter*, 27(4):59–62, October 1987b.

R. Collé and P. Karp. Measurement uncertainties: report of an international working group meeting. *Journal of Research of the National Bureau of Standards*, 92:243–244, May 1987. doi: 10.6028/jres.092.021.

D. R. Cox. Frequentist and Bayesian statistics: a critique (Keynote Address). In L. Lyons and M. K. Ünel, editors, *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pages 3–6, London, UK, 2006. Imperial College Press. ISBN 1-86094-649-6. doi: 10.1142/9781860948985_0001. Proceedings of PHYSTAT05, Oxford, UK, 12-15 September 2005.

A. C. Davison. *Statistical Models*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-73449-3. doi: 10.1017/CBO9780511815850.

A. Eagle. Probability and randomness. In A. Hájek and C. Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*, chapter 21. Oxford University Press, Oxford, UK, 2016. ISBN 978-0199607617. doi: 10.1093/oxfordhb/9780199607617.013.22.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

C. Eisenhart and R. Collé. Postscript to expression of the uncertainties of final results. In C. W. Solomon, R. D. Bograd, and W. R. Tilley, editors, *NBS Communications Manual for Scientific, Technical, and Public Information*, chapter Exhibit 2-E, pages 2–30–2–32. U.S. Dept. of Commerce, National Bureau of Standards, Gaithersburg, MD, 1980. URL https://catalog.hathitrust.org/Record/011389799. Chapter 15 of the NBS Administrative Manual.

P. Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341(1):91–137, 2005. doi: 10.1016/j.tcs.2005.03.054.

A. L. Ganesan, M. Rigby, A. Zammit-Mangion, A. J. Manning, R. G. Prinn, P. J. Fraser, C. M. Harth, K.-R. Kim, P. B. Krummel, S. Li, J. Mühle, S. J. O'Doherty, S. Park, P. K. Salameh, L. P. Steele, and R. F. Weiss. Characterization of uncertainties in atmospheric trace gas inversions using hierarchical Bayesian methods. *Atmospheric Chemistry and Physics*, 14(8):3855–3864, 2014. doi: 10.5194/acp-14-3855-2014.

L. J. Gleser. Assessing uncertainty in measurement. *Statistical Science*, 13(3): 277–290, August 1998.

M. Goldstein and J. V. Howard. A likelihood paradox. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):619–628, 1991. doi: 10.2307/2345591.

C. Hagwood. The calibration problem as an ill-posed inverse problem. *Journal of Statistical Planning and Inference*, 31(2):179–185, 1992. doi: 10.1016/0378-3758(92)90028-Q.

A. Hájek. Interpretations of probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, 2007. URL plato.stanford.edu/archives/win2007/entries/probability-interpret/.

J. Hodges, J. Viallon, P. J. Brewer, B. J. Drouin, V. Gorshelev, C. Janssen, S. Lee, A. Possolo, M.-A. H. Smith, J. Walden, and R. Wielgosz. Recommendation of a consensus value of the ozone absorption cross-section at 253.65 nm based on literature review. *Metrologia*, 53(3):034001, 2019. doi: 10.1088/1681-7575/ab0bdd.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, 186(1007):453–461, 1946. doi: 10.1098/rspa.1946.0056.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL www.bipm.org/en/publications/guides/gum.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.

Joint Committee for Guides in Metrology. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 3rd edition, 2012. URL https://jcgm.bipm.org/vim/en/. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 200:2012 (2017 version with minor corrections and informative annotations).

Joint Committee for Guides in Metrology. *Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2020. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM GUM-6:2020.

Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL www.bipm.org/en/publications/guides/gum.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

A. Koepke, T. Lafarge, A. Possolo, and B. Toman. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, 54(3): S34–S62, 2017. doi: 10.1088/1681-7575/aa6c0e.

A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Co., New York, NY, second edition, 1933. Translation edited by Nathan Morrison.

H. H. Ku. Expressions of imprecision, systematic error, and uncertainty associated with a reported value. In C. W. Solomon, R. D. Bograd, and W. R. Tilley, editors, *NBS Communications Manual for Scientific, Technical, and Public Information*, chapter Exhibit 2-E, pages 2–24–2–29. U.S. Dept. of

Commerce, National Bureau of Standards, Gaithersburg, MD, 1980. URL https://catalog.hathitrust.org/Record/011389799. Chapter 15 of the NBS Administrative Manual.

K. Landsman. Randomness? What Randomness? *Foundations of Physics*, 50 (2):61–104, 2020. doi: 10.1007/s10701-020-00318-8.

D. G. Mayo and A. Spanos. Error statistics. In P. S. Bandyopadhyay and M. R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 153–198. North-Holland, Amsterdam, 2011. ISBN 978-0-444-51862-0. doi: 10.1016/B978-0-444-51862-0.50005-8.

M. G. Morgan and M. Henrion. *Uncertainty — A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, NY, first paperback edition, 1992. 10th printing, 2007.

F. Mosteller and J. W. Tukey. *Data Analysis and Regression*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1977. ISBN 0-201-04854-X.

P. T. O'Conner. *Woe is I: The Grammarphobe's Guide to Better English in Plain English*. Riverhead Books, New York, NY, fourth edition, 2019. ISBN 978-0525533-054.

A. O'Hagan. Science, subjectivity and software (comment on articles by berger and by goldstein). *Bayesian Analysis*, 1(3):445–450, September 2006. doi: 10.1214/06-BA116G.

A. O'Hagan. The Bayesian Approach to Statistics. In T. Rudas, editor, *Handbook of Probability: Theory and Applications*, chapter 6. Sage Publications, Thousand Oaks, CA, 2008. ISBN 978-1-4129-2714-7. doi: 10.4135/9781452226620.n6.

A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 2015. doi: 10.6028/NIST.TN.1900. NIST Technical Note 1900.

A. Possolo and H. K. Iyer. Concepts and tools for the evaluation of measurement uncertainty. *Review of Scientific Instruments*, 88(1):011301, 2017. doi: 10.1063/1.4974274.

A. Possolo, A. Koepke, D. Newton, and M. R. Winchester. Decision tree for key comparisons. *Journal of Research of the National Institute of Standards and Technology*, 126:126007, 2021. doi: 10.6028/jres.126.007.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, NY, second edition, 2007. ISBN 978-0-387-71598-8.

C. D. Rodgers. *Inverse Methods for Atmospheric Sounding: Theory and Practice*, volume 2 of *Atmospheric, Oceanic, and Planetary Physics*. World Scientific, Singapore, 2000. ISBN 981-02-2740-X.

R. B. F. Schumacher. A dissenting position on uncertainties. *NCSL Newsletter*, 27(4):55–59, October 1987.

S. Stoudt, A. Pintar, and A. Possolo. Coverage intervals. *Journal of Research of the National Institute of Standards*, 126:126004, 2021. doi: 10.6028/jres.126.004.

M. Thompson and S. L. R. Ellison. Dark uncertainty. *Accreditation and Quality Assurance*, 16:483–487, October 2011. doi: 10.1007/s00769-011-0803-0.

B. Toman and A. Possolo. Model-based uncertainty analysis in inter-laboratory studies. In F. Pavese, M. Bär, A. B. Forbes, J. M. Linares, C. Perruchet, and N. F. Zhang, editors, *Advanced Mathematical and Computational Tools in Metrology and Testing: AMCTM VIII*, volume 78 of *Series on Advances in Mathematics for Applied Sciences*, pages 330–343. World Scientific Publishing Company, Singapore, 2009a. ISBN 981-283-951-8.

B. Toman and A. Possolo. Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 14:553–563, October 2009b. doi: 10.1007/s00769-009-0547-2.

B. Toman and A. Possolo. Erratum to: Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 15:653–654, 2010. doi: 10.1007/s00769-010-0707-4.

T. von Clarmann, D. A. Degenstein, N. J. Livesey, S. Bender, A. Braverman, A. Butz, S. Compernolle, R. Damadeo, S. Dueck, P. Eriksson, B. Funke, M. C. Johnson, Y. Kasai, A. Keppens, A. Kleinert, N. A. Kramarova, A. Laeng, B. Langerock, V. H. Payne, A. Rozanov, T. O. Sato, M. Schneider, P. Sheese, V. Sofieva, G. P. Stiller, C. von Savigny, and D. Zawada. Overview: Estimating and reporting uncertainties in remotely sensed atmospheric composition and temperature. *Atmospheric Measurement Techniques*, 13(8):4393–4436, 2020. doi: 10.5194/amt-13-4393-2020.

D. R. White. In pursuit of a fit-for-purpose uncertainty guide. *Metrologia*, 53: S107–S124, 2016. doi: 10.1088/0026-1394/53/4/S107.

K. Yoshino, D.E. Freeman, J.R. Esmond, and W.H. Parkinson. Absolute absorption cross-section measurements of ozone in the wavelength region 238-335 nm and the temperature dependence. *Planetary and Space Science*, 36(4): 395–398, 1988. doi: 10.1016/0032-0633(88)90127-4.