Statistical Science
1998, Vol. 13, No. 3, 277–290

Assessing Uncertainty in Measurement

Leon Jay Gleser

Abstract. In 1993 the International Organization for Standardization (ISO), in cooperation with several other international organizations, issued *Guide to the Expression of Uncertainty in Measurement* in order to establish, and standardize for international use, a set of general rules for evaluating and expressing uncertainty in measurement. The ISO recommendation has been of concern to many statisticians because it appears to combine frequentist performance measures and indices of subjective distributions in a way that neither frequentists nor Bayesians can fully endorse. The purpose of this review of the ISO *Guide* is to describe the essential recommendations made in the *Guide*, and then to show how these recommendations can be regarded as approximate solutions to certain frequentist and Bayesian inference problems. The framework thus provided will, hopefully, allow statisticians to develop improvements to the ISO recommendations (particularly in the approximations used), and also better communicate with the physical science researchers who will be following the ISO guidelines.

Key words and phrases: Accuracy, measurand, random errors, systematic bias, frequentist distributions, subjective distributions, combined measurement, propagation of errors, use of expert judgement, confidence intervals, credible intervals, estimation of risk.

1. INTRODUCTION

Measurements and the conclusions derived from them are the foundation of science and technology. No measurement can perfectly determine the value of the quantity being measured (the *measurand*). Imprecisions arising from flaws in the construction of the instrument, from operator error, from incorrect specification of environmental conditions or from failure to identify all factors determining the measurement output can lead the measurement to deviate from the measurand value. A measurement, together with current knowledge, can allow one to eliminate certain values as implausible, but there will be uncertainty about which of the remaining values is the correct one. A measure of the spread of the collection of values not rendered implausible by the measurement might be called the *uncertainty* of that measurement.

Scientists checking each other's conclusions about a certain measurand will compare measurements. They need to know what amount of disparity between measurements is acceptable. A measure of uncertainty should set limits for such differences, such that a difference outside of these limits suggests that different measurands are being measured. Thus, communication among scientists about the results of their measurements requires presentation of some index of the uncertainty of the measurement(s) involved.

In industry, measurements of processes are compared to requirements set by management or to quality standards set by the national standards laboratories. For example, the diameters of drilled openings in circuit boards are compared to values set by a manufacturer's specifications, whereas the viscosity of motor oil must meet standards decided upon and enforced by the oil industry. Instrumental calibration is typically done by using instruments on measurands whose values are "known" (standards). Knowledge of the amount of uncertainty in the measured values of these standards is required in order to determine if the instruments being calibrated need to be adjusted to remove systematic bias.

How should an index of uncertainty of measurement be defined and presented? For measurements that are (in theory) repeatable, with outcomes

Leon Jay Gleser is Professor, Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 (e-mail: ljg@bacchus.stat.pitt.edu).

whose deviations (errors) from the measurand appear random with mean zero, the traditional quantification of uncertainty for a single measurement has been some multiple of the standard deviation (or other suitable measure of spread) of the distribution of errors. If u is the uncertainty of a measurement m, the measurement and its uncertainty are often presented in the form

$$(1) \qquad m \pm u.$$

Unfortunately, in this notational convention the multiple of the error standard deviation used to calculate u varies widely from field to field, and even among researchers within a field. In physics, u is typically chosen to be the error standard deviation, so that (assuming that errors are approximately normally distributed) the interval (1) gives a 68% confidence interval for the value of the measurand. Other disciplines may set u to be 1.96 times the error standard deviation, so that (1) is a 95% confidence interval for the value of the measurand. At various times in the last century, prominent scientists (such as Churchill Eisenhart; see Eisenhart, 1968) have deplored the multiplicity of uncertainty indices and notational practices, and called for a standard way of presenting information about uncertainty in measurement.

As the processes treated in science and industry become more and more sophisticated, requiring extremely high degrees of accuracy, the simplistic model

$$(2) \qquad m = \mu + e, \quad e \text{ random with mean } 0,$$

has become less and less realistic because deviations from the measurand μ due to imprecisely determined contextual conditions (or imprecisions in the scientific models used) are now of such a magnitude that they cannot be ignored. These deviations tend to hold constant over repeated measurements (in a given context) and thus lead to systematic bias in measurement.

Consequently, the model

$$(2') \qquad m = \mu + b + e, \quad e \text{ random with mean } 0,$$

where b is the systematic bias, may be more appropriate. This is particularly true when b and the standard deviation of e are of the same order of magnitude. If b is known, then of course we can subtract b from m and create a new measurement that obeys model (2). It is much more often the case, however, that the value of b is not precisely known. Determining b may involve other measurements (with their own uncertainties) and also require scientific judgement. Scientific opinions are not infallible and can vary from expert to expert. If a range $[X-d, X+d]$ of values from b is known with near certainty to contain the correct value of b, then conservative methods for establishing a confidence interval for the value μ of the measurand have been proposed. For example, the method advocated by Eisenhart (1963) uses the midpoint X as a bias correction to m, and reports the interval

$$(3) \qquad (m - X) \pm [1.96(\text{st.dev. of } e) + d]$$

as a 95% confidence interval for the value of the measurand.

The method described by Eisenhart has been called "the (American) orthodox position." It has been used by the National Bureau of Standards (now called the National Institute of Standards and Technology) of the United States, and similar methods also have been recommended by the National Physical Laboratory of the United Kingdom. Apart from requiring specification of a range of values for the bias, this approach avoids reliance on individual scientific judgement and presents uncertainty about the measurand in frequentist terms, thus following the paradigm for statistical inference that has been the "orthodox" American and English approach to statistical inference in this century.

Indeed, the strong belief of supporters of the "orthodox position" that data-based measurement, being objective, was superior to "subjective" expert opinion, led them to make a strong distinction between random and systematic uncertainties, and to insist that these components of any overall uncertainty be separately reported. Many statisticians even felt that these components should not be combined at all. In response Eisenhart and Collé (1980) stated:

> Without depreciating the perils of shorthand expressions, there is often a need for an overall uncertainty statement which combines the imprecision and systematic uncertainty components. Arguments that it is incorrect from a theoretical point of view to combine the individual components in any fashion are not always practical. First, an approach which retains all details is not amenable for large compilations of results from numerous sources. And second, this approach shifts the burden of evaluating the uncertainties to users. Many users need a single uncertainty value resulting from the combination of all sources of inaccuracy. These users believe, and

> rightly so, that this overall estimate of inaccuracy can be most appropriately made by the person responsible for the measurement result.

The quantity "1.96(st.dev. of e) + d" can be regarded as such a combined uncertainty index for the measurement m. Even so, note that this index treats the random and systematic components of model (2′) differently.

The increase of international cooperation in industry and the merging of European commercial systems seems to have given new impetus to attempts to standardize the reporting of measurement uncertainty. Such standardization is seen as crucial to communication among national laboratories, and also to communication among and within multinational industries. In 1978 the International Bureau of Weights and Measures (BIPM) sent questionnaires to 32 national laboratories about their measurement methods and reporting of uncertainty. In the 21 replies received, the only unanimous agreement (not unexpectedly) was that the standard deviation be used for reporting "random uncertainty" (i.e., uncertainty arising from mean-zero random errors of measurement). Otherwise, there was a wide divergence of opinion, particularly on how to adjust for known sources of systematic bias.

In 1980 the BIPM convened a Working Group, consisting of representatives from 11 national standards laboratories. This Working Group produced a set of five rules for reporting uncertainty. The International Committee for Weights and Measures (CIPM) adopted these rules in 1981 and later reaffirmed support of the BIPM recommendation. In 1986, CIPM asked the International Organization for Standardization (ISO) to develop a detailed guide based on the BIPM recommendations "which . . . reflects the needs arising from the broad interests of industry and commerce." In 1993 the first edition of *Guide to the Expression of Uncertainty in Measurement* was published (ISO Technical Advisory Group, Working Group 3, 1993; hereafter, the *Guide*). In the same year, the National Institute of Standards and Technology published *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results* (Taylor and Kuyatt, 1993) that implements the BIPM/ISO approach.

If the BIPM/ISO approach is followed, there will be international standardization of presentations of uncertainty, at least by national laboratories; such standardization is certainly highly desirable. The methods advocated for expressing uncertainty in the *Guide*, however, seem to form a new paradigm for statistical inference that is neither completely frequentist nor completely Bayesian, and consequently lack a firm theoretical basis. Thus, there is concern among statisticians that the methods advocated by the *Guide* could prove to be misleading or inaccurate. Widespread acceptance of the paradigm advocated in the *Guide* within the physical science community also has the potential to increase confusion about statistical concepts and thus impede communication between statisticians and nonstatisticians.

The main points of the BIPM/ISO proposal for calculating and presenting uncertainty of measurement are summarized in Section 2. It is shown in Section 3 that the two types of uncertainty index mentioned in this proposal can each be posed as a solution to a certain frequentist inference problem. Bayesians will probably, and appropriately, object that the BIPM/ISO proposal does not go far enough in modeling and incorporating subjective (expert) opinion into the expression of uncertainty, and that indices of the posterior distribution of the measurand value should be reported in place of the BIPM/ISO uncertainty measures. In Section 4, however, it is shown that the BIPM/ISO measure of uncertainty provides an upper bound for the total Bayes risk of the measurement as an estimator of the measurand. Some generalizations of the results presented in Sections 3 and 4 are outlined in Section 5. Finally, the question whether a single universal measure of uncertainty can exist is briefly considered in Section 6.

2. THE BIPM/ISO PROPOSAL

2.1 The Basic Approach

The simple measurement context described in the Introduction is a special case of a more general measurement problem where the measurand μ is not necessarily directly measured, but rather is obtained as a function of p measured quantities $\theta_1, \ldots, \theta_p$, whose uncertainties are determined through statistical analysis of a series of observations (*Type A uncertainties*), and of r measurands $\lambda_1, \ldots, \lambda_r$, whose possible values and uncertainties (*Type B uncertainties*) are evaluated by other means, including expert judgement. That is,

$$(4) \qquad \mu = g(\theta_1, \ldots, \theta_p; \lambda_1, \ldots, \lambda_r) = g(\boldsymbol{\theta}; \boldsymbol{\lambda}),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)'$ and $g(\cdot; \cdot)$ is a known real-valued function.

As an example, we might wish to determine the power μ dissipated by a temperature-sensitive resistor which has resistance λ_1 at a specified temperature c and whose linear temperature coefficient of resistance is stated in references to be λ_2. A potential θ_1 is applied to the terminals of the resistor when the actual temperature is θ_2. Both the potential and the actual temperature are measured by instruments which have been statistically calibrated, whereas the resistance at temperature t and the linear temperature coefficient of resistance are determined by expert judgement (here, the use of references). In this case,

$$\mu = (\theta_1)^2[\lambda_1(1+\lambda_2(\theta_2 - c))]^{-1}.$$

Whereas the orthodox approach distinguished measurements by whether they had random or systematic errors, the BIPM/ISO approach characterizes measurements by the method through which their uncertainties are quantified. In the example just presented, the uncertainties of the measures of potential and temperature are quantified by calibration studies (that provide the user with estimated standard errors having specified degrees of freedom). Thus, these measurements have Type A uncertainties according to the BIPM/ISO classification; our notation reflects this classification.

To obtain uncertainties for Type B measurements, the BIPM/ISO *Guide* makes several recommendations. All of these recommendations have in common the establishment of a range of possible true values of the quantities in question, and the adoption of a probability distribution over this range. For a scalar parameter λ, the range of values might be $[X - d, X + d]$ and the chosen distribution uniform or triangular over this range. The distributional choice, however, serves primarily to justify formulas for the standard deviation in terms of the half-width d of the range of values. For example, this standard deviation is $d/3$ in the case of the uniform distribution. The standard deviation of the chosen distribution is taken to be the (*Type B*) *standard uncertainty* of the measurement of λ. If the distribution is assumed to be symmetric about X, or more generally has mean X, then X is taken to be the measurement of λ.

Let $Y_1, \ldots, Y_p$ be measurements of $\theta_1, \ldots, \theta_p$, respectively. Let $\sigma(Y_i)$ be the standard deviation of Y_i based on repeated use of the measuring instrument yielding Y_i, $i = 1, \ldots, p$. Let $s(Y_i)$ be an estimate of $\sigma(Y_i)$ based on repeated observations; where appropriate, let ν_i be the degrees of freedom of $s(Y_i)$. The (*Type A*) *standard uncertainty* of Y_i, denoted $u(Y_i)$, is defined to be $\sigma(Y_i)$ and is estimated by $s(Y_i)$, $i = 1, \ldots, p$. We have already indicated how measurements $X_1, \ldots, X_r$ of $\lambda_1, \ldots, \lambda_r$ and the (Type B) standard uncertainties $u(X_j)$ of these measurements are defined.

The BIPM/ISO *Guide* assumes that

$$(5) \qquad m = g(Y_1, \ldots, Y_p; X_1, \ldots, X_r) = g(\mathbf{Y}; \mathbf{X})$$

will be used as the measurement of μ. A key part of the BIPM/ISO recommendation is the method by which the *standard uncertainty* $u(m)$ of m is calculated.

Assuming existence and continuity of the first partial derivatives of $g(\cdot;\cdot)$, (5) is expanded in a first-order Taylor series about $(\boldsymbol{\theta}; \boldsymbol{\lambda})$:

$$(6) \qquad \begin{aligned} m &\approx \sum_{i=1}^{p} \frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \theta_i}(Y_i - \theta_i) \\ &\quad + \sum_{j=1}^{r} \frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_j}(X_j - \lambda_j) \end{aligned}$$

from which it follows that

$$(7) \qquad \begin{aligned} \operatorname{var}(m) &\approx \sum_{i=1}^{p} \left[\frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \theta_i}\right]^2 u^2(Y_i) \\ &\quad + \sum_{j=1}^{r} \left[\frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_j}\right]^2 u^2(X_j), \end{aligned}$$

assuming mutual statistical independence of the Y_i's and λ_j's. The approximation (7) will be reasonably accurate when the range of values for the Y_i's and λ_j's is sufficiently small that $g(\cdot;\cdot)$ is nearly linear over that range.

The standard uncertainty of m is (approximately) the square root of the right-hand side of (7) and can be estimated by

$$(8) \qquad \bar{u}(m) \approx \sqrt{\sum_{i=1}^{p} (c_i)^2 s^2(Y_i) + \sum_{j=1}^{r} (d_j)^2 u^2(X_j)},$$

where

$$c_i = \left[\frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \theta_i}\right]_{\boldsymbol{\theta}=\mathbf{Y},\, \boldsymbol{\lambda}=\mathbf{X}}, \quad d_j = \left[\frac{\partial g(\boldsymbol{\theta};\boldsymbol{\lambda})}{\partial \lambda_j}\right]_{\boldsymbol{\theta}=\mathbf{Y},\, \boldsymbol{\lambda}=\mathbf{X}},$$

and

$$Y = (Y_1, \ldots, Y_p)', \quad \mathbf{X} = (X_1, \ldots, X_r)'.$$

The Scatterthwaite–Welch formula is used to compute an "effective degrees of freedom" ν_{eff} for this estimate of uncertainty:

$$\nu_{\text{eff}} = \frac{[u(m)]^4}{\sum_{i=1}^{p} (c_i)^4 s^4(Y_i) \nu_i^{-1}}.$$

(Note that here, the Type B uncertainties have been assumed known, and thus have infinite degrees of freedom for purposes of computing approximate degrees of freedom. The *Guide* also suggests ways to assign degrees of freedom to Type B uncertainties when such uncertainties are only approximately known.)

The BIPM/ISO *Guide* calls (8) the *combined standard uncertainty* of the measurement m. An *extended combined uncertainty* for m is defined to be k times the standard uncertainty, where k is a constant (the *coverage factor*) chosen to make the interval

$$(9) \qquad m \pm k\bar{u}(m)$$

a $100(1-\alpha)\%$ "confidence" interval for the measurand μ. Choosing k to be the $100(1-\alpha/2)$ percentile of the t-distribution with degrees of freedom equal to ν_{eff} is said to serve as a reasonable approximation when the estimated Type A standard uncertainties have (approximate) chi-distributions.

Either the standard uncertainty or an extended uncertainty of a measurement m is acceptable as a quantitative report of uncertainty for the measurement. Note, however, that if m is later used as part of the determination of a measurand defined as a function of μ and other quantities, an extended uncertainty will have to be converted into a standard uncertainty for m in order to compute the combined standard uncertainty of the new measurement. For that reason, the *Guide* requires that k be separately specified whenever an extended measurement is reported.

In the context of the measurement model (2′) given in the Introduction, we can identify θ_1 with the sum $\mu + b$ of the measurand and bias, Y with the original measurement m, and λ_1 with the bias b. Then

$$\mu = g(\theta_1, \lambda_1) = \theta_1 - \lambda_1.$$

If the bias λ_1 is equally likely to take on any value in the interval $[X - d,\ X + d]$, and if Y has a normal distribution with known standard deviation (Type A standard uncertainty) $u(Y)$, then instead of the interval

$$(10) \qquad Y - X \pm [1.96\ u(Y) + d]$$

recommended by the "orthodox position" as a 95% confidence interval for the measurand μ, the BIPM/ISO *Guide* would recommend

$$(11) \qquad Y - X \pm 1.96\ [u^2(Y) + d^2/3]^{1/2}$$

as an approximate 95% "confidence" interval. When d is sufficiently large relative to $u(Y)$, the interval (11) contains the interval (10) even though (10) is known to be a ***conservative*** frequentist confidence interval for μ. This point raises conceptual issues about how the BIPM/ISO measures of uncertainty are to be interpreted.

2.2 Concepts

The *Guide*'s standard uncertainty of measurement is intended to be interpreted as the standard deviation of a probability distribution $h(\mu)$ of possible values for the measurand μ. The extended uncertainty is the half-width of an interval centered at m that yields a specified probability determined from the distribution $h(\cdot)$. The *Guide*'s interpretation of the distribution $h(\cdot)$, and of probabilities obtained from this distribution, is that such a distribution represents "degree of belief" about possible values of the measurand.

Although this interpretation seems to coincide with a Bayesian point of view, the distribution $h(\cdot)$ is not a posterior distribution for μ given the measurement m. Rather it is constructed partly from a frequentist distribution for the measurements Y (and their estimated uncertainties) and partly from a subjective (degree-of-belief) distribution for the quantities $\lambda_1, \ldots, \lambda_r$. To make these distributions comparable indications of degrees of belief, one must be prepared to argue that a distribution for a measurement $\mathbf{Y}$ of a measurand $\boldsymbol{\theta}$ can be reinterpreted as a distribution of degree of belief about the values of $\boldsymbol{\theta}$ based on the observation of $\mathbf{Y}$. This reinterpretation cannot be Bayesian in nature, for such a reinterpretation would be equivalent to asserting that the distributions of $\mathbf{Y}$ conditional on $\boldsymbol{\theta}$ and of $\boldsymbol{\theta}$ conditional on $\mathbf{Y}$ are always the same. Instead, certain comments made in the *Guide* seem to follow the spirit of R. A. Fisher's justification of fiducial probabilities. (Particularly revealing is the note at the end of page 45 that "It is assumed that probability is viewed as a measure of the degree of belief that an event will occur, implying that a systematic error may be treated in the same way as a random error and that ε_i represents either kind.") There are known conceptual problems with the general use of fiducial theory. Whether or not a revival of fiducial probabilities is the intention of the authors of the *Guide*, the theory underlying the interpretation of standard and extended uncertainties requires additional clarification.

2.3 Properties

Nevertheless, there are many attractive features of the BIPM/ISO method for determining standard uncertainties of measurements:

1. It uses a familiar measure of variability, the standard deviation, as its index of uncertainty, and uses a well-known method for propagating errors to combine component uncertainties.
2. It eliminates the necessity for distinguishing between random error and systematic bias, a dis-

tinction basic to the orthodox method, but (as argued in the *Guide*) often dependent on context. (For example, errors made in measurement due to individual foibles of the measurer can be viewed as systematic biases when only one individual is doing the measurements, but can be viewed as a component of variation when many individuals are responsible for measurements.) Even though the *Guide* distinguishes Type A and Type B uncertainties, this distinction is used only in deciding what information needs to be used to compute uncertainties. Once the standard uncertainty of a measurement is given, it is used in the same way regardless of whether it is Type A or Type B.

3. It is highly portable. The standard uncertainty $u(W)$ of a measurement W and its degrees of freedom (if $u(W)$ is estimated or approximated) can be used directly to compute the combined standard uncertainty and degrees of freedom for any measurement m composed from W. The standard uncertainty for m can in turn be used to compute the combined standard uncertainty for a new measurement m' composed from m, and so on. This is true regardless of whether $u(W)$ is a Type A or Type B uncertainty.
4. It does not require precise distributional assumptions. For measurands $\boldsymbol{\theta}$ with Type A uncertainties, it is not even necessary to assume the existence of (prior) distributions of degrees of belief.

Difficulties with the method are as follows:

(a) It appears to interpret standard deviations of measurement error distributions as if they were standard deviations of degrees of belief for distributions of Type A measurands.

(b) The accuracy of formula (6) depends on the curvature of $g(\cdot;\cdot)$, and may not be sufficiently precise for many applications. Using more terms in a Taylor expansion to account for the nonlinearity of $g(\cdot;\cdot)$ may fail to provide improvement, and also is more cumbersome.

The *Guide*'s extended uncertainties have desirable properties similar to those of the standard uncertainties, although (because of the need to specify the constant k) they are more complicated to use. The extended uncertainties also share all of the drawbacks of the standard uncertainties. In addition, they depend for their interpretation on distributional approximations whose applicability is not assured in every case. As already noted, the "confidence" associated with these extended uncertainties does not have a clear interpretation in terms of degrees of belief.

Fortunately, it is possible to give meaningful interpretation to the *Guide*'s standard and extended uncertainties in both frequentist and Bayesian terms, as seen in Sections 3 and 4, respectively.

3. A FREQUENTIST INTERPRETATION

Although the orthodox method takes a worst-case view of the quantities $\lambda_1, \dots, \lambda_r$, it is not a violation of the frequentist paradigm to seek instead to estimate $\mu = g(\boldsymbol{\theta};\boldsymbol{\lambda})$ under a weighted squared-error loss function:

$$(12)\qquad L(m,\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (m - g(\boldsymbol{\theta};\boldsymbol{\lambda}))^2 \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}.$$

(Indeed, weighted averages of distributions over "nuisance parameters" underlie the marginal likelihood approach (which is a frequentist approach) to statistical inference. See Kalbfleisch and Sprott (1970).) Here, $\pi(\boldsymbol{\lambda})$ is a nonnegative weight function defined on the possible values of $\boldsymbol{\lambda}$ that can represent degree of belief. It is assumed that both the integral of $\pi(\cdot)$ over r-dimensional space and the integral in (12) are finite. Because of this assumption, we can assume without loss of generality that $\pi(\cdot)$ is a probability density function. Consequently, even though we may regard some elements of $\boldsymbol{\lambda}$ as representing fixed unknown states, *mathematically* we can treat $\boldsymbol{\lambda}$ as if it were a random vector.

3.1 The Standard Uncertainty

In such a context, it seems reasonable to define the uncertainty of m to be the risk of m as an estimator of μ. Alternatively, so that m and its uncertainty $u(m)$ are measured in the same units, we can define the uncertainty of m to be the square root of the risk of m:

$$(13)\qquad u(m) = \{E[L(m,\boldsymbol{\theta})]\}^{1/2},$$

where the expected value is taken over the distribution of $\mathbf{Y}$. (Remember that $m = g(\mathbf{Y};\mathbf{X})$ and that $\mathbf{X}$ is a constant.) Strictly speaking, $u(m)$ is a function of $\boldsymbol{\theta}$, but this fact is suppressed in our notation so as to parallel the notation used in the *Guide*. The *Guide* would estimate $u(m)$ by substituting $\mathbf{Y}$ for $\boldsymbol{\theta}$.

If $f(\cdot \mid \theta)$ is the density of $\mathbf{Y}$, then

$$(14)\qquad u^2(m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (g(\mathbf{Y};\mathbf{X}) - g(\boldsymbol{\theta};\boldsymbol{\lambda}))^2 \cdot f(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\lambda})\, d\mathbf{Y}\, d\boldsymbol{\lambda}.$$

Consequently, $u^2(m)$ can be viewed as the expected value of the squared difference between $g(\mathbf{Y};\mathbf{X})$ and $g(\boldsymbol{\theta};\boldsymbol{\lambda})$ over the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\lambda}$. Note that in this joint distribution, $\mathbf{Y}$ and $\boldsymbol{\lambda}$ are statistically independent.

Now, note that $m = g(\mathbf{Y};\mathbf{X})$ can be regarded as a random variable resulting from one of a family of transformations $g(\cdot;\mathbf{X})$ of $\mathbf{Y}$ indexed by $\mathbf{X}$. In Section 2 it was (implicitly) assumed that $\mathbf{Y}$ is an unbiased estimator of $\boldsymbol{\theta}$. Hence, if $g(\cdot;\mathbf{X})$ were a linear transformation of $\mathbf{Y}$, it would follow that $E(m) = E[g(\mathbf{Y};\mathbf{X})] = g(\boldsymbol{\theta};\mathbf{X})$. The extent to which $E(m)$ differs from $g(\boldsymbol{\theta},\mathbf{X})$ depends on the amount of nonlinearity in the transformation $g(\cdot;\mathbf{X})$.

Similarly, $\mu = g(\boldsymbol{\theta};\boldsymbol{\lambda})$ can be viewed as a random variable resulting from one of a family of transformations $g(\boldsymbol{\theta};\cdot)$ of $\boldsymbol{\lambda}$ indexed by $\boldsymbol{\theta}$. Assume that $E(\boldsymbol{\lambda}) = \mathbf{X}$, as would be the case if each component of $\boldsymbol{\lambda}$ were uniformly distributed over an interval whose midpoint is the corresponding component of $\mathbf{X}$. If $g(\boldsymbol{\theta};\cdot)$ were a linear transformation, then $E(\mu) = g(\boldsymbol{\theta};\mathbf{X})$. Thus, when $g(\mathbf{a};\mathbf{b})$ is linear in both $\mathbf{a}$ and $\mathbf{b}$, $E(m)$ equals $E(\mu)$.

NOTE. Observe one essential difference between measurements with Type A uncertainties and measurements or states with Type B uncertainties. Measurements $\mathbf{Y}$ with Type A uncertainties vary randomly about the fixed unknown value $\boldsymbol{\theta}$ of the measurand, whereas values of the measurand $\boldsymbol{\lambda}$ for measures of states having Type B uncertainties vary randomly around the observed measurement $\mathbf{X}$.

Regardless of whether or not $g(\cdot;\cdot)$ is linear, it is straightforward to show that

$$\begin{aligned} u^2(m) &= E(m-\mu)^2 \\ &= \mathrm{Var}(m) + (E(m)-E(\mu))^2 + \mathrm{Var}(\mu), \end{aligned} \tag{15}$$

noting in (14) that $\mathbf{Y}$ and $\boldsymbol{\lambda}$ are statistically independent.

When $g(\cdot;\cdot)$ is sufficiently close to being linear in both arguments, then $E(m)$ is approximately equal to $E(\mu)$ and use of one-term Taylor expansions to approximate the variances on the right-hand side of (15), estimating $\boldsymbol{\theta}$ by $\mathbf{Y}$, yields the square of the right-hand side of (8). This provides some frequentist justification for the *Guide*'s recommended standard uncertainty of m.

3.2 Extended Uncertainty

The *Guide*'s extended uncertainty may be defined in frequentist terms by requiring the interval $m \pm ku(m)$, whose half-width is the extended uncertainty $ku(m)$, to have minimum weighted coverage probability satisfying

$$\begin{aligned} \inf_{\boldsymbol{\theta}} \biggl[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P\{-ku(m) \le m - \mu \le ku(m)\} \\ \cdot \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \biggr] \\ \ge 1-\alpha, \end{aligned} \tag{16}$$

where $\pi(\boldsymbol{\lambda})$ is a nonnegative weight function defined over the values of $\boldsymbol{\lambda}$. As already noted, $\pi(\cdot)$ without loss of generality can be assumed to be a probability density function and thus $\boldsymbol{\lambda}$, although conceptually fixed, can be mathematically treated as if it were a random vector. (Recall again that $\mu = g(\boldsymbol{\theta};\boldsymbol{\lambda})$ and that $m = g(\mathbf{Y};\mathbf{X})$, where $\mathbf{X}$ is the assumed mean vector of the distribution of $\boldsymbol{\lambda}$.)

NOTE. If the standard uncertainties of the elements of the vector $\mathbf{Y}$ have to be estimated, then $u(m)$ in (16) will involve these (random) estimates, and the probability being integrated on the left-hand-side of (16) will concern these random estimates of uncertainty as well as the random vector $\mathbf{Y}$.

Let

$$I(\mathbf{Y},\boldsymbol{\lambda}) = \begin{cases} 1, & \text{if } |\,g(\mathbf{Y};\mathbf{X}) - g(\boldsymbol{\theta};\boldsymbol{\lambda})\,| \le ku(m), \\ 0, & \text{otherwise,} \end{cases}$$

be the indicator of the event $\{-ku(m) \le m - \mu \le ku(m)\}$, and let $f(\mathbf{Y} \mid \boldsymbol{\theta})$ be the density function of $\mathbf{Y}$. (If $u(m)$ is estimated from observations other than $\mathbf{Y}$, think of $f(\cdot \mid \boldsymbol{\theta})$ as being the joint density of these observations as well as $\mathbf{Y}$.) Define

$$C(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I(\mathbf{Y},\boldsymbol{\lambda}) f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\lambda})\, d\mathbf{Y}\, d\boldsymbol{\lambda}. \tag{17}$$

In terms of $C(\boldsymbol{\theta})$, requirement (16) becomes

$$\inf_{\boldsymbol{\theta}} C(\boldsymbol{\theta}) \ge 1-\alpha. \tag{18}$$

Note that $C(\boldsymbol{\theta})$ is a coverage probability defined from the joint distribution of $\mathbf{Y}$ (and any other data used to estimate $u(m)$) and $\boldsymbol{\lambda}$.

The *Guide*'s recommendations for determining k involve a great many approximations whose applicability is not always clear. First, the use of normal or t tables to find the multiplicative constant k seems to require that $m - \mu = g(\mathbf{Y};\mathbf{X}) - g(\boldsymbol{\theta};\boldsymbol{\lambda})$ is approximately normally distributed (for fixed $\mathbf{X}$ and $\boldsymbol{\theta}$). The *Guide* appeals to the central limit theorem and Taylor series approximations involving (weighted) sums of the elements of $\mathbf{Y}$ and $\boldsymbol{\lambda}$ to support such approximations. Note, however, that even when $\mathbf{Y}$ has an exact multivariate normal distribution and $g(\mathbf{a};\mathbf{b})$ is linear in both $\mathbf{a}$ and $\mathbf{b}$, the convolution $g(\mathbf{Y};\mathbf{X}) - g(\boldsymbol{\theta};\boldsymbol{\lambda})$ need not have a normal distribution if the dimension r of $\boldsymbol{\lambda}$ is 1 or 2 (and the components of $\boldsymbol{\lambda}$ are independent uniform random variables). The *Guide* does cite references that indicate that convolutions of i.i.d. uniform (or of symmetrical unimodal) distributions share with the normal distribution the property that approximately 90–95% of the total probability lies within two standard deviations of the mean, but the components of $\mathbf{Y}$ and $\boldsymbol{\lambda}$ are not, in general, identically distributed, and the components of $\mathbf{Y}$ need not have symmetrical distributions. Further, to use this result about convolutions, the standard uncertainty $u(m)$ must be known (or else approximated so accurately that the error has small effect).

If $u(m)$ must be estimated, the Scatterthwaite–Welch approximate t-calculations recommended by the *Guide* have doubtful validity. If the possibility that $\mathbf{Y}$ is based on data series that are not normally distributed is entertained, then the squared standard uncertainties of the elements of $\mathbf{Y}$ estimated from this data need not have chi-squared distributions and need not be independent of (or even uncorrelated with) the elements of $\mathbf{Y}$. Consequently, even if $g(\mathbf{a};\mathbf{b})$ is approximately linear in $\mathbf{a}$, the pivotal quantity $(m-\mu)/u(m)$ need not have a distribution that can be approximated by a t-distribution. One other point to note is that the weights in the Taylor series expansion (6) depend on the unknown parameter $\boldsymbol{\theta}$ and thus are estimated in (8) by substituting $\mathbf{Y}$ for $\boldsymbol{\theta}$. Consequently, both the numerator and denominator of the pivotal quantity

$$\frac{(m-\mu)}{u(m)} = \frac{g(\mathbf{Y};\mathbf{X})-\mu}{u(m)}$$

depend on $\mathbf{Y}$.

Although the left-hand side of (16) is the formal definition of the (weighted) confidence of the interval $m \pm ku(m)$, recent work on confidence interval estimation has tended to concentrate on the coverage probability $C(\boldsymbol{\theta})$ at the "true" value of $\boldsymbol{\theta}$, rather than the minimal coverage probability. For example, bootstrap confidence intervals are constructed by choosing the interval endpoints as a function of the data so as to make a bootstrapped *estimate* of $C(\boldsymbol{\theta})$ equal the desired coverage probability $1-\alpha$. [See DiCiccio and Efron (1996) and discussion following.] Choosing k as a function of the data in this way may provide a superior measure of uncertainty free of the ad hoc approximations that are of concern above. [See Taylor and Kuyatt (1993) for a similar recommendation.]

3.3 Commentary

Viewed as the risk of a frequentist inference procedure, (14) most closely resembles squared error prediction. That is, a statistic $m = g(\mathbf{Y})$ is used to predict the value of an independent random variable μ under squared error loss. This is done, according to familiar theory, by choosing the statistic m to estimate the mean of μ. In the present case, the mean of μ is

$$G(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{\theta};\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\,d\boldsymbol{\lambda}. \tag{19}$$

Realizing that estimation of $G(\boldsymbol{\theta})$ rather than $g(\boldsymbol{\theta};\mathbf{X})$ is required makes $m = g(\mathbf{Y};\mathbf{X})$ less intuitively attractive. It might be preferable to use $G(\mathbf{Y})$ instead, particularly if $G(\boldsymbol{\theta})$ is approximately linear in $\boldsymbol{\theta}$.

In place of a formula that estimates the variance of μ based on the approximate linearity of $g(\mathbf{a};\mathbf{b})$ as a function of $\mathbf{b}$, it might be preferable to compute

$$\begin{aligned}\tau^2(\boldsymbol{\theta}) &= \operatorname{var}(\mu)\\ &= \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^2(\boldsymbol{\theta};\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\,d\boldsymbol{\lambda}\right] - [G(\boldsymbol{\theta})]^2,\end{aligned} \tag{20}$$

and estimate the variance of μ by $\tau^2(\mathbf{Y})$. The integrals (19) and (20) can be computed by a variety of numerical methods, including Monte Carlo sampling; in some cases, exact analytical expressions can be derived. Note that if $G(\mathbf{Y})$ is used in place of $g(\mathbf{Y};\mathbf{X})$ as an estimate of μ, then the estimated squared uncertainty of $G(\mathbf{Y})$ is

$$\begin{aligned}u^2(G(\mathbf{Y})) &= \operatorname{var}(G(\mathbf{Y})) + \operatorname{var}(g(\mathbf{Y};\boldsymbol{\lambda}))\\ &\quad + (E[G(\mathbf{Y})] - G(\boldsymbol{\theta}))^2,\end{aligned} \tag{21}$$

which is approximately equal to $\operatorname{var}(G(\mathbf{Y})) + \boldsymbol{\tau}^{\mathbf{2}}(\mathbf{Y})$ when $G(\boldsymbol{\theta})$ is approximately linear in $\boldsymbol{\theta}$.

It might be argued that requiring one to specify the distribution $\pi(\boldsymbol{\lambda})$ reduces the portability of the measurement $G(\mathbf{Y})$ and its associated uncertainty estimate (21). This would be true if the distributions used were general. In fact, the *Guide* discusses primarily uniform and triangular distributions for those measurands λ_j evaluated by expert judgement (where the experts provide the endpoints of the range of possible values), and normal distributions for other quantities having Type B uncertainty of measurement. (These are usually combined measurements constructed from statistical series of measurements and expert judgement.) Consequently, if $\boldsymbol{\lambda}$ is assumed to have independent components, no more information is needed to compute (19) and (20) than is needed to compute approximations using Taylor's expansions.

The coverage probability $C(\boldsymbol{\theta})$ defined in (17) can be thought of as giving the probability that m predicts μ with error no greater than $ku(m)$, where both m and μ are random quantities defined as transformations of the observable random vector $\mathbf{Y}$ and of the unobservable random vector $\boldsymbol{\lambda}$, respectively. The interval $m \pm ku(m)$ satisfying (18) can thus be regarded as a $100(1-\alpha)\%$ prediction interval for μ.

4. A BAYESIAN INTERPRETATION

Bayesians will point to the *Guide*'s probabilistic modeling of subjective opinion concerning the elements of $\boldsymbol{\lambda}$ and question why the same is not done for the elements of $\boldsymbol{\theta}$. The *Guide* does not answer that question, and many statisticians believe that the answer may be political: a compromise made between those who see no distinction between frequentist and subjective probabilities and those who have been taught to distrust the use of subjective opinion in scientific inference.

Yet, it is possible that the failure to probabilistically model subjective opinion about the elements of $\boldsymbol{\theta}$ can be given a robust Bayesian explanation. To obtain useful results, the authors of the *Guide* are willing to incorporate probability distributions for quantities λ that cannot be evaluated by statistical analysis of observations, yet also are sufficiently accurately known that any possible deviations from their assumed values make only a small (but not negligible) contribution to deviations in the measurand $\mu = g(\boldsymbol{\lambda}; \boldsymbol{\theta})$ of primary interest. In such circumstances, errors in assigning probability distributions to the values of such quantities do not materially influence the measure of uncertainty produced, particularly when (as in the *Guide*'s proposal) only a few summary indices of such distributions are actually utilized in the analysis. On the other hand, the fact that series of observations are used to estimate the elements of $\boldsymbol{\theta}$ and their uncertainties suggests that less is known about the values of $\boldsymbol{\theta}$, or that variation in the values of $\boldsymbol{\theta}$ produces relatively greater variation in the values of the measurand μ. In this situation, errors in probabilistically modeling opinion about the values of $\boldsymbol{\theta}$ would be of considerable concern. Because no prior distribution can adequately model uncertainty in a way satisfactory to everyone, conclusions that are nearly independent of the choice of distribution for $\boldsymbol{\theta}$ are desirable.

A thoroughgoing Bayesian analysis of each measurement problem would be likely to involve extensive computation (particularly if the dimensions p and r of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are large), including analysis of the sensitivity to the prior distribution of the estimate of $\boldsymbol{\theta}$. Although considerable progress has been made recently in Bayesian computation, the existing methodology requires programs and expertise not readily available to the typical scientist, engineer or technician. The best reporting of uncertainty about the measurand would, of course, be the posterior distribution of that measurand. Such a distribution is not as portable as the *Guide*'s measure of uncertainty. (For example, it would be difficult to attach the accurate form of the posterior distribution to a measurand sent for comparative measurement to another scientist, whereas a single number such as $u(m)$ is easily attached.) An index of spread, such as the half-width of a highest posterior density (HPD) credible interval, could be reported in a single use, but the full posterior distribution will be needed if evaluation of μ and its uncertainty is needed as part of a future measurement process.

What, then, can be said in a Bayesian sense about the *Guide*'s proposed measure(s) of uncertainty? As before, let $f(\mathbf{Y} \mid \boldsymbol{\theta})$ be the density of the measurement vector $\mathbf{Y}$, where we continue to assume that this distribution does not depend on $\boldsymbol{\lambda}$. Let $\Gamma(\boldsymbol{\theta})$ be the (prior) density of $\boldsymbol{\theta}$, let $\pi(\boldsymbol{\lambda})$ be the (prior) density of $\boldsymbol{\lambda}$ and let $\mathbf{X}$ be the (prior) mean vector of $\pi(\boldsymbol{\lambda})$. Recall that

$$\mu = g(\boldsymbol{\theta}; \boldsymbol{\lambda}), \quad m = g(\mathbf{Y}; \mathbf{X}),$$

$$G(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{\theta}; \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}.$$

Let

$$\Gamma(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) \Gamma(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{Y} \mid \boldsymbol{\theta}^*) \Gamma(\boldsymbol{\theta}^*)\, d\boldsymbol{\theta}^*}$$

be the posterior density of $\boldsymbol{\theta}$. Then the posterior mean of the measurand μ can easily be shown to be

$$\mu(\mathbf{Y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G(\boldsymbol{\theta}) \Gamma(\boldsymbol{\theta} \mid \mathbf{Y})\, d\boldsymbol{\theta}. \tag{22}$$

Further, note that $E[\mu \mid \boldsymbol{\theta}] = G(\boldsymbol{\theta})$, and that

$$\begin{aligned} \tau^2(\boldsymbol{\theta}) &= \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^2(\boldsymbol{\theta}; \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda})\, d\boldsymbol{\lambda} \right] - [G(\boldsymbol{\theta})]^2 \\ &= \operatorname{Var}(\mu \mid \boldsymbol{\theta}) \end{aligned}$$

is the part of the variation of the measurand $\mu = g(\boldsymbol{\theta}; \boldsymbol{\lambda})$ that is attributable to the variation of $\boldsymbol{\lambda}$. The posterior variance of μ is then

$$\begin{aligned} v(\mathbf{Y}) = &\left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\tau^2(\boldsymbol{\theta}) + G^2(\boldsymbol{\theta})] \Gamma(\boldsymbol{\theta} \mid \mathbf{Y})\, d\boldsymbol{\theta} \right\} \\ &- [\mu(\mathbf{Y})]^2. \end{aligned} \tag{23}$$

A Bayesian might choose to report the posterior mean $\mu(\mathbf{Y})$ as the measurement, in place of m, and to report $[v(\mathbf{Y})]^{1/2}$ as the measure of uncertainty in place of $u(m)$. However, unless the posterior distribution of $G(\boldsymbol{\theta})$ given $\mathbf{Y}$ is approximately normal (or at least is symmetric unimodal), there is no necessary proportional relationship between $[v(\mathbf{Y})]^{1/2}$

and the half-width of an HPD credible interval for μ (which would be a good Bayesian candidate for an extended standard uncertainty measure for μ).

To what extent is the *Guide*'s measurement m and standard uncertainty $u(m)$ an approximation to these Bayesian measures? Let

$$\begin{aligned} H(\boldsymbol{\theta}) &= E[m \mid \boldsymbol{\theta}] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{Y};\mathbf{X}) f(\mathbf{Y} \mid \boldsymbol{\theta})\, d\mathbf{Y}, \\ u^2(\boldsymbol{\theta}) &= \mathrm{Var}[m \mid \boldsymbol{\theta}] \\ &= \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^2(\mathbf{Y};\mathbf{X}) f(\mathbf{Y} \mid \boldsymbol{\theta})\, d\mathbf{Y} \right\} \\ &\quad - [H(\boldsymbol{\theta})]^2. \end{aligned}$$

It can be shown (see Parzen, 1960, page 387) that the best linear (in m) Bayes estimator of $\mu = g(\boldsymbol{\theta}; \boldsymbol{\lambda})$, in the sense of least expected squared error over the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$, is

$$\mu_{\mathrm{Linear}}(\mathbf{Y}) = E_\Gamma[G(\boldsymbol{\theta})] + r[m - E_\Gamma[H(\boldsymbol{\theta})]],$$

where

$$r = \frac{\mathrm{Cov}_\Gamma(\mu, m)}{\mathrm{Var}_\Gamma(m)}$$

and the subscript Γ on the expected value, covariance and variance indicates that these computations are made relative to the prior distribution, $\Gamma(\boldsymbol{\theta})$. The total (expected posterior) Bayes risk of $\mu_{\mathrm{Linear}}(\mathbf{Y})$ is

$$\begin{aligned} &\text{total Bayes risk of } \mu_{\mathrm{Linear}}(\mathbf{Y}) \\ &\quad = r^2 \mathrm{Var}_\Gamma(m) + \mathrm{Var}_\Gamma(\mu). \end{aligned} \tag{24}$$

Because the linear Bayes estimator based on m is generally not equal to the Bayes estimator (22), the total Bayes risk for estimating μ is less than or equal to the total Bayes risk (24) of $\mu_{\mathrm{Linear}}(\mathbf{Y})$. Note that the posterior Bayes risk $v(\mathbf{Y})$ of $\mu(\mathbf{Y})$ is an estimator of the total Bayes risk, in the sense that the expected value of $v(\mathbf{Y})$ over the marginal distribution of $\mathbf{Y}$ is the total Bayes risk.

Suppose that

$$\frac{\mathrm{Var}_\Gamma[m - \mu]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} \approx 0. \tag{25}$$

Then it is shown in the Appendix that $r \approx 1$. Thus

$$\begin{aligned} &\text{total Bayes risk of } \mu_{\mathrm{Linear}}(\mathbf{Y}) \\ &\quad \approx \mathrm{var}_\Gamma(m) + \mathrm{var}_\Gamma(\mu) \approx u^2(m). \end{aligned} \tag{26}$$

Thus when (25) holds, the total Bayes risk of the linear (in m) Bayes estimator of μ under squared error loss is approximately equal to $u^2(m)$. This somewhat weakly justifies $u(m)$ as a measure of uncertainty from a Bayesian viewpoint.

Because $u^2(m)$ approximates the total Bayes risk of $\mu_{\mathrm{Linear}}(\mathbf{Y})$, which is in turn greater than or equal to the total Bayes risk for the Bayes estimator of μ, it follows that $u^2(m)$ is possibly biased as an estimator of the total Bayes risk of the Bayes estimator of μ. It is not possible from this analysis, however, to say that $u^2(m)$ is always greater than or equal to $v(\mathbf{Y})$.

The assumption made in (25) corresponds to asserting that the variability of m around μ is small relative to the variability of $G(\boldsymbol{\theta})$, which measures that part of the prior uncertainty of μ due to uncertainty about the values of $\boldsymbol{\theta}$. This assumption agrees with the scenario mentioned at the beginning of this section that attempted to justify why uncertainty in the values of $\boldsymbol{\lambda}$ was probabilistically modeled in the *Guide*, but uncertainty in the values of $\boldsymbol{\theta}$ was not.

NOTE 1. As remarked in Section 3, the function $G(\cdot)$ can be computed from the known distribution $\pi(\boldsymbol{\lambda})$. Note that the measurement $\mathbf{Y}$ actually is used to estimate $G(\boldsymbol{\theta})$ rather than $\mu = g(\boldsymbol{\theta}, \boldsymbol{\lambda})$. The total Bayes risk for estimating μ is the sum of the averaged (over $\Gamma(\theta)$) mean square error of m as an estimator of $G(\boldsymbol{\theta})$ and the averaged conditional variance of μ given $\boldsymbol{\theta}$. Other functions of $\mathbf{Y}$ (such as $G(\mathbf{Y})$) may be preferable to m as estimators of $G(\boldsymbol{\theta})$. The approximations shown above apply also to these alternative functions, with appropriate changes of definitions [e.g., of $H(\boldsymbol{\theta})$].

NOTE 2. The estimates of the uncertainties of the components of $\mathbf{Y}$ have not explicitly appeared in the discussion of this section. To include them, it is only necessary to extend the definition of $\mathbf{Y}$ to include both the measurement of $\boldsymbol{\theta}$ and any estimates of the component uncertainties of this measurement. The steps in the above discussion, and the main conclusions, remain unaffected by this change. This is, in part, because we have not devoted attention to how a Bayesian might want to estimate the total Bayes risk of $\mu_{\mathrm{Linear}}(\mathbf{Y})$. The most appropriate estimate of this risk, to a Bayesian, would be the posterior Bayes risk of $\mu_{\mathrm{Linear}}(\mathbf{Y})$, and this quantity would depend upon the estimates of the uncertainties of the components of $\mathbf{Y}$.

5. GENERALIZATIONS

5.1 Correlated Measurement Errors

In Section 2, the assumption of mutual statistical independence of the components of $\mathbf{Y}$ was utilized. This assumption was employed only in deriving (7) and (8). The intent in presenting these formulas was to present the ideas underlying the *Guide*'s proposal without getting involved in too much notation and detail.

The *Guide* presents generalizations of its variance propagation formulas to the case where (some of) the measurement errors of some of the components of $\mathbf{Y}$ are correlated. In this case, estimates of the covariances between the components of $\mathbf{Y}$, obtained from series of observations, are required. The *Guide* also shows how to use one-step Taylor series approximations to determine the covariance between two measurements constructed from the same basic measurements $\mathbf{Y}$ or measured under similar environments (common components of $\boldsymbol{\lambda}$), or both.

Because covariances between measurements of a vector of measurands are needed in uncertainty calculations, the definition of uncertainty in terms of standard deviations (or variances) lacks needed generality. The obvious generalization is to define the standard uncertainty of a vector of dependent measurements to be the (estimated) ensemble of standard deviations and correlations of the differences of those measurements from their corresponding measurands; or, equivalently, by the estimated variance–covariance matrix of such differences. But now, rather than being required only to report one measure of uncertainty for each measurement, one would need to report that uncertainty plus any needed covariances or correlations with other measurements. The portability and convenience of the *Guide*'s proposal is thus not as great as might have, at first glance, appeared to be the case.

5.2 Dependence between *Y* and λ

Throughout Sections 2 through 4, it was assumed that the measurands $\boldsymbol{\lambda}$ were statistically independent of the errors of measurement in $\mathbf{Y}$. This assumption is also implicitly made in the *Guide* because the authors regard measurement errors as being those involved in the raw act of measurement, and not in any adjustments made to try to correct for the fact that the true state $\boldsymbol{\lambda}$ of the environment or physical system did not have the value $\mathbf{X}$ assumed for it. This point of view is made explicit in Sections 5.2.4 and 5.2.5 of the *Guide*, where it is noted that two measurements can be correlated because their values are corrected for common deviations from an hypothesized state $\mathbf{X}$. In this case, the *Guide* recommends that the measurements be redefined to have the values they had before such corrections were made, and that the common state be included separately as a Type B measurand.

One can perhaps imagine situations in which measurement errors involved in the raw act of measurement have a distribution that depends on a state of the environment that is accounted for by $\boldsymbol{\lambda}$. As long as the raw errors of measurement conditional on $\boldsymbol{\lambda}$ have expected value 0, however, the raw errors of measurement will be uncorrelated with $\boldsymbol{\lambda}$, and the analysis given in Sections 2 and 3 would be unaffected. The formulas shown for the exact Bayes estimator and its posterior Bayes risk in Section 4 are, of course, invalidated when the raw measurement errors in $\mathbf{Y}$ are statistically dependent on $\boldsymbol{\lambda}$, but the conclusions reached about the interpretation of the *Guide*'s proposals from a Bayesian perspective remain unaffected.

6. THE MANY TYPES OF UNCERTAINTY

A scientist is interested in a measurand μ that is a function of certain measurands $\theta_1, \dots, \theta_p$ and states $\lambda_1, \dots, \lambda_r$. The scientist measures $\theta_1, \dots, \theta_p$ using averages $Y_1, \dots, Y_r$ of series of observations on these measurands. The scientist's observations also provide estimates of the standard uncertainties of $Y_1, \dots, Y_p$. The scientist uses $Y_1, \dots, Y_p$ and assumed values $X_1, \dots, X_r$ of the states to construct a measurement m of μ. The scientist's findings are to be sent to four colleagues A, B, C and D. Should the same standard uncertainty index be reported to all of them?

Colleague A wants to use the value of this measurand in a forthcoming article. For this colleague, the *Guide*'s standard uncertainty $u(m)$ for the measurement m and μ is appropriate.

Colleague B will repeat the scientist's measurement in a different environmental context. The measurand μ is assumed to be contextually invariant, so that the measurement m^* of this colleague is statistically independent of m and has the same uncertainty. Consequently, $2^{1/2}u(m)$ is the standard deviation of the difference $m - m^*$ and can be used to determine whether the two measurements are actually of the same measurand (value). Further, $m - m^* \pm t[2^{1/2}u(m)]$, for an appropriate value of t, provides an approximate confidence interval for the difference of the two measurand values (if any). Because comparison of measurements is a common use for an index of uncertainty, terminology is needed to distinguish the uncertainty $u(m)$ of the measurand from the uncertainty $2^{1/2}u(m)$ of the measurement process. (See also Ku, 1990.)

Colleague C also wishes to compare his or her measurement m' of μ with the scientist's measurement m. This colleague's measurement, however, will be taken in the same environment as the scientist's. Because the environments are the same for the two measurements, the portion of the variation of m (and m') due to uncertainty concerning the states $\lambda_1, \dots, \lambda_r$ cancels out in the difference

$m - m'$. Thus $u(m)$, which incorporates this component of variation, cannot by itself be used to determine the standard deviation of $m - m'$. In this situation, $u(m)$ should reflect only the variation in m due to measurement errors in the Y_i's.

Finally, Colleague D wishes to use the scientist's measurement m to help in the measurement of a new measurand ν that is a function of μ and of other measurands and states considered by the scientist. Again, the scientist cannot help this colleague by simply reporting the standard uncertainty $u(m)$, but needs also to report standard uncertainties for those measurands and states used by both the scientist and this colleague so that the covariance between the two measurements (m and the measurement of ν) can be determined.

By now, the point being made should be clear: *What we report as measures of uncertainty depends on the use that is intended for such measures*. Thus there can be no single universal index of uncertainty, although in each case a different combined uncertainty index can be used as an overall summary of measurement quality. Given the multiple purposes for which a measurement may be used, the best advice, made also by the *Guide*, is to report enough information about the probability distributions of measures of the basic quantities (the θ_i's and λ_j's) involved in the construction of the measurand μ to allow users to compute indices of uncertainty of their choice. Crude summaries based on standard deviations and correlations may be enough for most measurands, but for expensive, hard-to-replicate measurements, more detailed records need to be provided. One of the dangers of proposals such as that made in the *Guide* is that they may encourage a minimal reportage that can cause important statistical information to be discarded.

7. CONCLUSION

Viewing the *Guide*'s proposal for computing an index of uncertainty concerning the values of a measurand as a crude approximation to either a frequentist or a Bayesian problem may hopefully help reconcile statisticians to an approach that at first glance seems to be entirely ad hoc. If all of the *Guide*'s recommendations are followed, there will be clearer and more self-consistent reports of accuracy or uncertainty in science and technology, so that a researcher in one field will not have to learn the reporting conventions of another field in order to make sense of measurement summaries in the journals of that field. The *Guide* also encourages reports of uncertainties of component measurements, so that more detailed summaries of results may be reported than is presently the case. When used for the same purpose in each case, the measure of uncertainty proposed by the *Guide* is portable from study to study, permitting (at a more crude level) the type of updating of statistical information that is one of the great advantages of the Bayesian paradigm.

Perhaps the major problem with (and criticism of) the *Guide*'s proposal is that the assumptions underlying its approximations can be violated, and even if such assumptions hold approximately, it is not clear how accurate these approximations are in practical contexts. Although there is a considerable literature on "large-sample approximations," it is remarkable how little is known about distributions arising from sums of independent, but not identically distributed, random summands, or more generally about the accuracy of distributional approximations based on Taylor series expansions. In consequence, the definition of a standard uncertainty can more easily be supported (in terms of solutions to standard frequentist or Bayesian problems) than can the definition of an expanded uncertainty, which requires approximate normality (or at least approximate distributional symmetry) of the measurement to justify use of the standard deviation in constructing confidence or credible regions. For many researchers, however, a measure of uncertainty is useful only if it provides one with the ability to make confidence or posterior probability statements about measurands, so that research is needed to indicate the situations in which the *Guide*'s approximations do, or do not, provide confidence or credible regions of acceptable accuracy.

One can regard the reporting of measurements and their uncertainty as a decision theory problem. Following Lu and Berger (1989), the action to be taken can be a pair (m, u), where m is a point estimate of a measurand μ, and where u (or u^2) is an estimate of the risk of m under some specified loss function. Alternatively (Kiefer, 1977; Hwang and Brown, 1991), one can consider a pair in which the first element is an interval estimate of the measurand and the second element is an estimate of the coverage probability of this interval. These papers, and many of the references they cite, emphasize methods for determining the admissibility of the second component of the pair (i.e., of u). The results and methods of analysis that they present may be useful in approaching the problem of how to report uncertainty. Such analyses, however, can have little impact on the measurement field unless any improvements made on the ISO's recommendations are nonnegligible and demonstrable, software is provided to implement such improvements, and the improvements are in performance characteristics of interest and concern to measurement specialists.

Whatever statisticians may think of the *Guide*'s proposals for reporting accuracy of measurement, this proposal is certain to influence the thinking of physical science researchers about statistical concepts. To react by saying "It's not right and you shouldn't use it" is likely to be futile, for the advantages of the ISO proposal are compelling to individuals not interested in statistical theory. Instead, what is needed is research that reveals the limitations of the ISO approach, improvements that extend its applicability with a minimum of extra conceptualization and computation and statisticians able to communicate such results to nonspecialists.

APPENDIX

PROOF THAT (25) IMPLIES THAT $r \approx 1$. It is well known (see DeGroot, 1986, page 242, Exercise 25) that

$$\begin{aligned}\mathrm{Cov}_\Gamma(m,\mu) &= \mathrm{Cov}_\Gamma(E[m \mid \boldsymbol{\theta}], E[\mu \mid \boldsymbol{\theta}]) \\ &= \mathrm{Cov}_\Gamma(H(\boldsymbol{\theta}), G(\boldsymbol{\theta})) \\ &= \mathrm{Cov}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta}), G(\boldsymbol{\theta})] \\ &\quad + \mathrm{Var}_\Gamma[G(\boldsymbol{\theta})].\end{aligned} \tag{A.1}$$

Using the formula $\mathrm{Var}(z) = E[\mathrm{Var}(z \mid w)] + \mathrm{Var}[E[z \mid w]]$ repeatedly (see DeGroot, 1986, Exercise 10, page 225),

$$\begin{aligned}&\mathrm{Var}_\Gamma(m-\mu) \\ &\quad = \mathrm{Var}_\Gamma(m) + \mathrm{Var}_\Gamma(\mu) - 2\,\mathrm{Cov}_\Gamma(m,\mu) \\ &\quad = E_\Gamma(\mathrm{Var}[m|\boldsymbol{\theta}]) + \mathrm{Var}_\Gamma(E[m \mid \boldsymbol{\theta}]) \\ &\qquad + E_\Gamma(\mathrm{Var}[u \mid \boldsymbol{\theta}]) + \mathrm{Var}_\Gamma(E[\mu \mid \boldsymbol{\theta}]) \\ &\qquad - 2\,\mathrm{Cov}_\Gamma(m,\mu) \\ &\quad = E_\Gamma[u^2(\boldsymbol{\theta})] + \mathrm{Var}_\Gamma[H(\boldsymbol{\theta})] + E_\Gamma[\tau^2(\boldsymbol{\theta})] \\ &\qquad + \mathrm{Var}_\Gamma[G(\boldsymbol{\theta})] - 2\,\mathrm{Cov}_\Gamma[H(\boldsymbol{\theta}), G(\boldsymbol{\theta})] \\ &\quad = E_\Gamma[u^2(\boldsymbol{\theta})] + E_\Gamma[\tau^2(\boldsymbol{\theta})] \\ &\qquad + \mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})],\end{aligned} \tag{A.2}$$

and also

$$\mathrm{Var}_\Gamma[m] = E_\Gamma[u^2(\boldsymbol{\theta})] + \mathrm{Var}_\Gamma[H(\boldsymbol{\theta})]. \tag{A.3}$$

Note that

$$\begin{aligned}&\mathrm{Var}_\Gamma[H(\boldsymbol{\theta})] \\ &\quad = \mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta}) + G(\boldsymbol{\theta})] \\ &\quad = \mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})] + \mathrm{Var}_\Gamma[G(\boldsymbol{\theta})] \\ &\qquad + 2\,\mathrm{Cov}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta}), G(\boldsymbol{\theta})].\end{aligned} \tag{A.4}$$

Because the terms $E_\Gamma[u^2(\boldsymbol{\theta})]$, $E_\Gamma[\tau^2(\boldsymbol{\theta})]$ and $\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]$ are all nonnegative, it follows from (A.2) and (25) that

$$\begin{aligned}\frac{\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} &\approx 0, \\ \frac{E_\Gamma(u^2(\boldsymbol{\theta}))}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} &\approx 0.\end{aligned} \tag{A.5}$$

Thus, from (A.1) and (A.5),

$$\begin{aligned}&\left|\frac{\mathrm{Cov}_\Gamma(m,\mu)}{\mathrm{Var}_\Gamma(G(\boldsymbol{\theta}))} - 1\right| \\ &\quad = \frac{|\,\mathrm{Cov}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta}), G(\boldsymbol{\theta})]\,|}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} \\ &\quad \le \frac{\{\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]\,\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]\}^{1/2}}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} \\ &\quad = \left\{\frac{\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]}\right\}^{1/2} \\ &\quad \approx 0,\end{aligned} \tag{A.6}$$

and it follows similarly from (A.3), (A.4) and (A.5) that

$$\begin{aligned}&\left|\frac{\mathrm{Var}_\Gamma[m]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} - 1\right| \\ &\quad \le \frac{E_\Gamma[u^2(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} + \frac{\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} \\ &\qquad + 2\left\{\frac{\mathrm{Var}_\Gamma[H(\boldsymbol{\theta}) - G(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]}\right\}^{1/2} \\ &\quad \approx 0.\end{aligned}$$

Consequently,

$$\begin{aligned}r &= \frac{\mathrm{Cov}_\Gamma(m,\mu)}{\mathrm{Var}_\Gamma(m)} = \frac{\mathrm{Cov}_\Gamma(m,\mu)/\mathrm{Var}_\Gamma(G(\boldsymbol{\theta})]}{\mathrm{Var}_\Gamma(m)/\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})]} \\ &\approx \frac{1}{1} = 1,\end{aligned}$$

as was to be shown. □

Note that, for r to be approximately 1, it is sufficient for (A.5) to hold, rather than (25). That is, to obtain the result $r \approx 1$, we do not need to assume that $E_\Gamma[\tau^2(\boldsymbol{\theta})]/\mathrm{Var}_\Gamma[G(\boldsymbol{\theta})] \approx 0$.

ACKNOWLEDGMENTS

The author has benefitted from discussion of the BIPM/ISO proposal with Dr. Keith R. Eberhardt while the author was a visitor at the National Institute of Standards. An earlier and less complete draft of this paper was presented at a conference, "Multiple Decision Theory and Related Topics," held in honor of Shanti S. Gupta, June 8–12, 1995. Work on the present paper was partially supported by NSF Grant DMS-95-04924.

REFERENCES

DeGroot, M. H. (1986). *Probability and Statistics*, 2nd ed. Addison-Wesley, Reading, MA.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statist. Sci.* **11** 189–229.

Eisenhart, C. (1963). Realistic evaluation of the precision and accuracy of instrument calibration. *Journal of Research of the National Bureau of Standards* **67C** 161–187. [Reprinted, with corrections, in (1969) *Precision Measurement and Calibration: Statistical Concepts and Procedures*, National Bureau of Standards Special Publication 300 (H. H. Ku, ed.) **1** 21–48. U.S. Government Printing Office, Washington, DC.]

Eisenhart, C. (1968). Expression of the uncertainties of final results. *Science* **160** 1201–1204.

Eisenhart, C. and Collé, R. (1980). Postscript. *National Bureau of Standards Manual for Scientific, Technical and Public Information* November 2-30 to 2-32. [Reprinted in Eisenhart, C., Ku, H. H. and Collé, R. (1983). *Expression of the Uncertainties of Final Measurement Results*. National Bureau of Standards Special Publication 644, U.S. Government Printing Office, Washington, DC.]

Hwang, J. T. and Brown, L. D. (1991). Estimated confidence under the validity constraint. *Ann. Statist.* **19** 1964–1977.

ISO Technical Advisory Group, Working Group 3 (1993). *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, Geneva.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

Ku, H. H. (1990). *Uncertainty and Accuracy in Physical Measurements*. National Institute of Standards and Technology Special Publication 805, U.S. Government Printing Office, Washington, DC.

Lu, K. L. and Berger, J. O. (1989). Estimation of normal means: frequentist estimation of loss. *Ann. Statist.* **17** 890–906.

Parzen, E. (1960). *Modern Probability Theory and Its Applications*. Wiley, New York.

Taylor, B. N. and Kuyatt, C. E. (1993). Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST Note 1297, Physics Laboratory, National Institute of Standards and Technology, Gaithersburg, MD.